

Cours et exercices corrigés

2e et 3e cycles/Master • Écoles d'ingénieurs

TRAITEMENT NUMÉRIQUE DU SIGNAL

Théorie et pratique

8e édition

Préface de Pierre Aigrain

Maurice Bellanger



TRAITEMENT NUMÉRIQUE DU SIGNAL

Théorie et pratique

TRAITEMENT NUMÉRIQUE DU SIGNAL

Théorie et pratique

Maurice Bellanger
Professeur au Conservatoire National des Arts et Métiers à Paris
Membre de l'Académie des Technologies

Préface de
Pierre Aigrain
Membre de l'Académie des Sciences
Ancien Secrétaire d'État à la Recherche

8e édition



Illustration de couverture : *Digital Vision*

ISBN 2 10 050162 3

PRÉFACE

Les révolutions techniques les plus importantes et les plus riches de conséquences ne sont pas toujours celles qui sont les plus visibles pour l'utilisateur final du produit. Les méthodes modernes de traitement numérique du signal entrent dans la catégorie des révolutions techniques aux conséquences encore insuffisamment perçues et qui ne font pas la première page des journaux.

Il est intéressant d'ailleurs de réfléchir quelques instants à la manière dont de telles techniques voient le jour. Le traitement par le calcul numérique d'un signal au sens le plus large du terme n'est certes pas en soi une idée nouvelle. Lorsque Kepler tirait les lois du mouvement des planètes des séries d'observations de son beau-père Tycho Brahé, c'est à un véritable traitement numérique du signal qu'il se livrait, le signal en l'occurrence étant constitué par les séries temporelles des observations de positions de Tycho Brahé. Mais ce n'est que dans le courant de ces toutes dernières décennies que le traitement numérique du signal est devenu une discipline en soi : c'est que la nouveauté tient à ce que l'on peut maintenant procéder, en temps réel, au traitement de signaux électriques, et ceci par des méthodes numériques.

Pour que cette évolution soit possible, il fallait que des progrès techniques, dans de nombreux domaines, voient progressivement le jour, et tout d'abord, bien sûr, la possibilité d'acquérir, sous forme de signal électrique, des informations à traiter. Cela impliquait le développement progressif de tout ce qu'il est parfois convenu d'appeler les capteurs d'informations, lesquels peuvent aller, dans leur complexité, de la simple jauge de contrainte (mais il a fallu de nombreuses et difficiles recherches de physique des solides pour la rendre possible) au radar.

Il fallait aussi que se développent, avec les prodigieux progrès de la micro-électronique, les outils technologiques permettant de réaliser, aux cadences extrêmement élevées qu'implique le traitement en temps réel, des opérations arithmétiques que les premiers ordinateurs (l'ENIAC n'a que 40 ans) ne pouvaient réaliser qu'en plusieurs heures souvent interrompues de plusieurs pannes, et que nous trouvons aujourd'hui tout à fait naturel de voir exécutées par un micro-processeur de quelques grammes consommant seulement quelques milliwatts, et dont le temps moyen entre pannes dépasse dix ans.

Il fallait enfin que les méthodes de programmation, c'est-à-dire d'utilisation optimisée de ces outils nouveaux, aient pu progresser, car quelles que soient les immenses capacités de calcul des micro-processeurs modernes, il n'est pas indifférent de ne pas gaspiller ces possibilités en opérations inutiles. L'invention des algorithmes

de transformée de Fourier rapide est un des exemples les plus frappants de cette importance des méthodes de programmation. Cette convergence des progrès techniques dans des domaines aussi différents relevant pour les uns de la physique, pour beaucoup de l'électronique, pour d'autres des mathématiques, n'a pas été accidentelle. Dans une certaine mesure, chacun des progrès a suscité le besoin nouveau auquel un nouveau progrès dans un autre domaine permettait de répondre. Il serait sans doute utile, du point de vue de l'histoire et de l'épistémologie des Sciences et des Techniques d'entreprendre un jour une étude approfondie de ce cas.

Car les conséquences en sont d'ores et déjà considérables. Sans doute le traitement analogique de signaux électriques a-t-il précédé le traitement numérique et sans doute continuera-t-il à occuper une place importante dans certaines applications, mais les avantages du traitement numérique qui tiennent en deux mots « précision et fiabilité » ont seuls rendu possibles certaines réalisations et qui débordent de loin les secteurs de l'électronique et des télécommunications dans lesquels ces techniques ont vu le jour. Pour n'en citer qu'une, la tomodensitographie par rayon X, les « scanners » sont basés sur l'application d'un théorème dû à Radon et connu depuis 1917. Seules les évolutions que nous avons mentionnées plus haut ont rendu possible la réalisation pratique de ce nouvel outil de diagnostic médical. Il y a gros à parier que les techniques de traitement numérique du signal trouveront demain leur place dans des produits de plus en plus variés, y compris les produits utilisés par le grand public qui, tout en bénéficiant des avantages de prix, de performance et de fiabilité que ces techniques rendent possibles, ne se rendra pas toujours compte de la prodigieuse imbrication de recherche, de technique et d'invention que suppose ce progrès. Cette évolution a d'ailleurs déjà commencé dans le cas des récepteurs de télévision.

Mais lorsque se produisent de telles révolutions techniques, une autre difficulté se rencontre presque inévitablement. C'est celle de la formation des utilisateurs à ce qui est non seulement un nouvel outil, mais souvent un nouveau mode de pensée. Cette étape de la formation peut devenir, si l'on n'y prend garde, un goulot d'étranglement dans l'introduction de nouvelles techniques. C'est pourquoi l'ouvrage de M. Bellanger, dont le point de départ est un enseignement donné depuis plusieurs années à l'École Nationale Supérieure des Télécommunications et à l'Institut Supérieur d'Électronique de Paris, constitue un événement dont il convient de souligner l'importance. Ouvrage didactique, accompagné d'exercices, contenant plusieurs programmes, que certains pourront souvent utiliser tel quel, il contribuera sans aucun doute à accélérer encore une évolution désirable et nécessaire.

P. AIGRAIN
1981

TABLE DES MATIÈRES

CONTENTS

AVANT-PROPOS

L'innovation impose à l'ingénieur une mise à jour permanente de ses connaissances et une bonne information sur le potentiel offert par les techniques nouvelles, découvertes et mises au point dans les laboratoires de recherche. En traitement du signal, les techniques numériques apportent des possibilités prodigieuses : la conception rigoureuse des systèmes, une grande reproductibilité des équipements, une grande stabilité de leurs caractéristiques en exploitation et une remarquable facilité de supervision. Cependant, ces techniques présentent un certain degré d'abstraction et leur application aux cas concrets requiert un ensemble de connaissances théoriques, jugées souvent plus familières ou plus facilement accessibles au chercheur qu'à l'ingénieur, et qui peuvent représenter un obstacle à leur utilisation. L'ambition du présent ouvrage est de vaincre cet obstacle et de faciliter l'accès aux techniques numériques en faisant la liaison entre la théorie et la pratique, et en mettant à la portée de l'ingénieur les résultats les plus utiles dans ce domaine.

La base de cet ouvrage est un enseignement donné dans des écoles d'ingénieurs, d'abord l'École Nationale Supérieure des Télécommunications et l'Institut Supérieur d'Électronique de Paris, puis Supélec et le CNAM. Il s'adresse donc d'abord aux ingénieurs. L'auteur s'est efforcé d'y faire une présentation claire et concise des principales techniques de traitement numérique, de comparer leurs mérites et de donner les résultats les plus utiles sous une forme directement exploitable aussi bien pour la conception des systèmes que pour une évaluation rapide dans le cadre de l'élaboration d'un projet en temps limité. Les développements théoriques ont été réduits à ce qui est nécessaire pour une bonne compréhension et une application correcte des résultats. Le lecteur trouvera dans les références bibliographiques les compléments qu'il pourrait souhaiter. A la fin de chaque chapitre, quelques exercices, souvent tirés de cas concrets, permettent de tester l'assimilation de la matière du chapitre et de se familiariser avec son utilisation. Pour ces exercices, des éléments de réponse et des indications ont été regroupés en fin d'ouvrage. Il convient également de signaler que des efforts ont été faits pour introduire une terminologie française, qu'il serait souhaitable de compléter et généraliser afin de donner à notre langue sa place à part entière dans le domaine.

Cet ouvrage s'adresse également aux chercheurs à qui il peut apporter, en plus d'un ensemble de résultats utiles, des indications pour l'orientation de leurs travaux, en faisant clairement apparaître les contraintes de la réalité technique. Il contient de plus un certain nombre de résultats provenant des travaux de

recherche de l'auteur et de ses collaborateurs. En effet, pour établir le dialogue avec les chercheurs et être en état de faire bénéficier la technique de leurs découvertes dans les délais les plus brefs, l'ingénieur doit s'intégrer à la communauté scientifique et apporter sa propre contribution à la recherche ; par ses contacts permanents avec les aspects concrets, il peut non seulement évaluer et conforter les résultats obtenus par les chercheurs mais encore ouvrir de nouvelles voies.

Par rapport aux précédentes, cette huitième édition apporte des compléments, des parties nouvelles et des simplifications. Les compléments portent sur des résultats partiels nouveaux introduits dans différents chapitres, des expressions simplifiées et des éclaircissements. Les parties nouvelles se situent principalement aux chapitres 11 et 13. En effet, il est apparu intéressant de montrer les connexions entre les filtres QMF et les ondelettes et que les ondelettes élémentaires résultent simplement d'un calcul de filtre avec des contraintes particulières. Quant à l'analyse et modélisation des signaux, ces fonctions s'intègrent de plus en plus dans les systèmes. Le traitement du signal se généralisant avec l'électronique, le domaine d'application est de plus en plus vaste et il a semblé judicieux de se limiter à quelques illustrations des méthodes de base.

Il faut souligner que les travaux sur lesquels est basé le présent ouvrage ont été à l'origine menés en collaboration et avec le soutien du Centre National d'Études des Télécommunications, à qui l'auteur tient à exprimer sa reconnaissance. Il tient également à exprimer sa profonde gratitude à Monsieur J. DAGUET, Directeur technique à la Société Télécommunications Radioélectriques et Téléphoniques pour avoir guidé ses travaux avec une grande clairvoyance et les avoir efficacement stimulés pendant de nombreuses années. L'auteur adresse aussi ses vifs remerciements à l'ensemble de ses collaborateurs pour leurs contributions et pour l'assistance constante qu'ils ont apportée.

INTRODUCTION

Le signal est le support de l'information émise par une source et destinée à un récepteur; c'est le véhicule de l'intelligence dans les systèmes. Il transporte les ordres dans les équipements de contrôle et de télécommande, il achemine sur les réseaux l'information, la parole ou l'image. Il est particulièrement fragile et doit être manipulé avec beaucoup de soins. Le traitement qu'il subit a pour but d'extraire des informations, de modifier le message qu'il transporte ou de l'adapter aux moyens de transmission; c'est là qu'interviennent les techniques numériques. En effet, si l'on imagine de substituer au signal un ensemble de nombres qui représentent sa grandeur ou amplitude à des instants convenablement choisis, le traitement, même dans sa forme la plus élaborée, se ramène à une séquence d'opérations logiques et arithmétiques sur cet ensemble de nombres, associées à des mises en mémoire.

La conversion du signal continu analogique en un signal numérique est réalisée par des capteurs qui opèrent sur des enregistrements ou directement dans les équipements qui produisent ou reçoivent le signal. Les opérations qui suivent cette conversion sont réalisées par des calculateurs numériques agencés ou programmés pour effectuer l'enchaînement des opérations définissant le traitement.

Avant d'introduire le contenu des différents chapitres du présent ouvrage, il convient de donner une définition précise du traitement considéré.

Le traitement numérique du signal désigne l'ensemble des opérations, calculs arithmétiques et manipulations de nombres, qui sont effectués sur un signal à traiter, représenté par une suite ou un ensemble de nombres, en vue de fournir une autre suite ou un autre ensemble de nombres, qui représentent le signal traité. Les fonctions les plus variées sont réalisables de cette manière, comme l'analyse spectrale, le filtrage linéaire ou non linéaire, le transcodage, la modulation, la détection, l'estimation et l'extraction de paramètres. Les machines utilisées sont des calculateurs numériques.

Les systèmes correspondant à ce traitement obéissent aux lois des systèmes discrets. Les nombres sur lesquels il porte peuvent dans certains cas être issus d'un processus discret. Cependant, ils représentent souvent l'amplitude des échantillons d'un signal continu et dans ce cas, le calculateur prend place derrière un dispositif convertisseur analogique-numérique et éventuellement devant un convertisseur numérique-analogique. Dans la conception de tels systèmes et l'étude de leur fonctionnement, la numérisation du signal revêt une importance fondamentale et les opérations d'échantillonnage et de codage doivent être analysées dans leur principe et leurs conséquences. La théorie des distributions constitue une approche concise, simple et efficace pour cette analyse. Après un certain nombre de rappels

sur l'analyse de Fourier, les distributions et la représentation des signaux, le chapitre premier rassemble les résultats les plus importants et les plus utiles sur l'échantillonnage et le codage d'un signal.

L'essor du traitement numérique date de la découverte d'algorithmes de calcul rapide de la Transformée de Fourier Discrète. En effet, cette transformation est à la base de l'étude des systèmes discrets et elle constitue dans ce domaine numérique l'équivalent de la Transformation de Fourier dans le domaine analogique, c'est le moyen de passage de l'espace des temps discret à l'espace des fréquences discret. Elle s'introduit naturellement dans une analyse spectrale avec un pas de fréquence diviseur de la fréquence d'échantillonnage des signaux à analyser.

Les algorithmes de calcul rapide apportent des gains tels qu'ils permettent de faire les opérations en temps réel dans de nombreuses applications pourvu que certaines conditions élémentaires soient remplies. Ainsi, la Transformation de Fourier Discrète constitue non seulement un outil de base dans la détermination des caractéristiques du traitement et dans l'étude de ses incidences sur le signal, mais de plus, elle donne lieu à la réalisation d'équipements toutes les fois qu'une analyse de spectre intervient, par exemple, dans les systèmes comportant des bancs de filtres ou quand, par la puissance de ses algorithmes, elle conduit à une approche avantageuse pour un circuit de filtrage. Les chapitres 2 et 3 lui sont consacrés ; ils donnent d'une part une présentation des propriétés élémentaires et du mécanisme des algorithmes de calcul rapide et de leurs applications, et d'autre part, un ensemble de variantes associées aux situations pratiques. En tant que système, le calculateur de Transformée de Fourier Discrète est un système linéaire discret, invariant dans le temps.

Une grande partie du présent ouvrage est consacrée à l'étude des systèmes linéaires discrets invariants dans le temps à une dimension, qui sont facilement accessibles et très utiles. Les systèmes à plusieurs dimensions et en particulier à deux et trois dimensions connaissent un grand développement ; ils sont appliqués par exemple aux images ; cependant, leurs propriétés se déduisent en général de celles des systèmes à une dimension dont ils ne sont souvent que des extensions simplifiées. Les systèmes non linéaires ou variables dans le temps, soit contiennent un sous-ensemble important qui présente les propriétés de linéarité et invariance temporelle, soit peuvent s'analyser avec les mêmes techniques que les systèmes ayant ces propriétés.

La linéarité et l'invariance temporelle entraînent l'existence d'une relation de convolution qui régit le fonctionnement du système, ou filtre, ayant ces propriétés. Cette relation de convolution est définie à partir de la réponse du système au signal élémentaire que représente une impulsion, la réponse impulsionnelle, par une intégrale dans le cas des signaux analogiques. Ainsi, si $x(t)$ désigne le signal à filtrer, $h(t)$ la réponse impulsionnelle du filtre, le signal filtré $y(t)$ est donné par l'équation :

$$y(t) = \int_{-\infty}^{\infty} h(\tau)\, x(t-\tau)\, d\tau$$

Dans ces conditions, une telle relation qui pourtant traduit directement le fonctionnement réel du filtre, offre un intérêt pratique limité. En effet, d'une part il n'est pas très aisé de déterminer la réponse impulsionnelle à partir des critères qui définissent l'opération de filtrage envisagée et d'autre part une équation comportant une intégrale ne permet pas facilement de reconnaître et vérifier le comportement du filtre. La conception est beaucoup plus facile à aborder dans le domaine des fréquences car la transformation de Laplace ou la transformation de Fourier permettent d'accéder à un plan transformé où les relations de convolution du plan amplitude-temps deviennent de simples produits de fonctions. A la réponse impulsionnelle, la transformation de Fourier fait correspondre la réponse en fréquence du système, et le filtrage se ramène au produit de cette réponse en fréquence par la transformée de Fourier, ou spectre, du signal à filtrer.

Dans les systèmes numériques, qui sont du type discret, la convolution se traduit par une sommation. Le filtre est défini par une suite de nombres qui constitue sa réponse impulsionnelle. Ainsi, si la suite à filtrer s'écrit $x(n)$, la suite filtrée $y(n)$ s'exprime par la sommation suivante, où n et m sont des entiers :

$$y(n) = \sum_m h(m)\, x(n-m)$$

Deux cas se présentent alors. Ou bien la sommation porte sur un nombre fini de termes, c'est-à-dire que les $h(m)$ sont nuls sauf pour un nombre fini de valeurs de la variable entière m. Le filtre est dit à réponse impulsionnelle finie; en faisant allusion à sa réalisation, on le désigne encore par non récursif car il ne nécessite pas de boucle de réaction de la sortie sur l'entrée dans sa mise en œuvre. Il est à mémoire finie, puisqu'il ne garde le souvenir d'un signal élémentaire, une impulsion par exemple, que pendant une durée limitée. Les nombres $h(m)$ sont appelés les coefficients du filtre, qu'ils définissent complètement. Ils peuvent se calculer d'une manière directe très simple, par exemple en faisant le développement en série de Fourier de la réponse en fréquence à réaliser. Ce type de filtre présente des caractéristiques originales très intéressantes; par exemple, la possibilité d'une réponse rigoureusement linéaire en phase, c'est-à-dire d'un temps de propagation de groupe constant; les signaux dont les composantes se trouvent dans la bande passante du filtre ne sont pas déformés à la traversée de ce filtre. Cette possibilité est exploitée dans les systèmes de transmission de données ou en analyse spectrale par exemple.

Ou bien la sommation porte sur un nombre infini de termes, les $h(m)$ ont une infinité de valeurs non nulles; le filtre est dit à réponse impulsionnelle infinie ou encore de type récursif, car il faut réaliser sa mémoire par une boucle de réaction de la sortie sur l'entrée. Son fonctionnement est régi par une équation selon laquelle un élément de la suite de sortie $y(n)$ est calculée par la sommation pondérée d'un certain nombre d'éléments de la suite d'entrée $x(n)$ et d'un certain nombre d'éléments de la suite de sortie précédents. Par exemple, si L et K sont des entiers, le fonctionnement du filtre peut être défini par l'équation suivante :

$$y(n) = \sum_{l=0}^{L} a_l x(n-l) - \sum_{k=1}^{K} b_k y(n-k)$$

Les $a_l (l = 0, 1, \ldots, L)$ et $b_k (k = 1, 2, \ldots, K)$ sont les coefficients. Comme pour les filtres analogiques, l'étude de ce type de filtre ne se fait pas en général simplement de manière directe ; il est nécessaire de passer par un plan transformé. La transformation de Laplace ou la transformation de Fourier pourraient être utilisées. Cependant, il existe une transformation beaucoup mieux adaptée, la transformation en Z, qui est l'équivalent pour les systèmes discrets. Un filtre est caractérisé par sa fonction de transfert en Z, désignée généralement par H(Z), et qui fait intervenir les coefficients par l'équation suivante :

$$H(Z) = \frac{\sum_{l=0}^{L} a_l Z^{-l}}{1 + \sum_{k=1}^{K} b_k Z^{-k}}$$

Pour obtenir la réponse en fréquence du filtre, il suffit de remplacer dans H(Z) la variable Z par l'expression suivante où f désigne la variable fréquence et T la période d'échantillonnage des signaux :

$$Z = e^{j2\pi f T}$$

Dans cette opération, à l'axe imaginaire, dans le plan de Laplace, correspond le cercle de rayon unité centré à l'origine dans le plan de la variable Z. Il apparaît clairement que la réponse en fréquence du filtre défini par H(Z) est une fonction périodique ayant pour période la fréquence d'échantillonnage. Une autre représentation de la fonction H(Z) est utile pour la conception des filtres et l'étude d'un certain nombre de propriétés, celle qui fait apparaître les racines du numérateur appelées zéros du filtre, $Z_l (l = 1, 2, \ldots, L)$ et les racines du dénominateur appelées pôles, $P_k (k = 1, 2, \ldots, K)$:

$$H(Z) = a_0 \frac{\prod_{l=1}^{L} (1 - Z_l Z^{-1})}{\prod_{k=1}^{K} (1 - P_k Z^{-1})}$$

Le terme a_0 est un facteur d'échelle qui définit le gain du filtre. La condition de stabilité du filtre s'exprime très simplement par la contrainte suivante : tous les pôles doivent être à l'intérieur du cercle unité. La position des pôles et des zéros par rapport au cercle unité, permet une appréciation très simple et très utilisée des caractéristiques du filtre.

Un ensemble de quatre chapitres est consacré à l'étude des caractéristiques de ces filtres numériques. Le chapitre IV présente les propriétés des systèmes linéaires discrets invariants dans le temps, rappelle les propriétés principales de la

transformation en Z et donne les éléments nécessaires à l'étude des filtres. Le chapitre V traite des filtres à réponse impulsionnelle finie : leurs propriétés sont étudiées, les techniques de calcul des coefficients sont décrites ainsi que les structures de réalisation. Les filtres à réponse impulsionnelle infinie étant généralement réalisés par une mise en cascade de cellules élémentaires du premier et second ordre ; le chapitre VI décrit ces cellules et leurs propriétés, ce qui d'une part facilite considérablement l'approche de ce type de système et d'autre part fournit un ensemble de résultats très utiles dans la pratique. Le chapitre VII donne les méthodes de calcul des coefficients des filtres à réponse impulsionnelle infinie et traite les problèmes apportés par la réalisation, avec les limitations qu'elle implique et leurs conséquences, en particulier le bruit de calcul.

Les filtres à réponse impulsionnelle infinie ayant des propriétés comparables à celles des filtres analogiques continus, il est naturel d'envisager pour leur réalisation des structures du même type que celles qui sont couramment employées en filtrage analogique. C'est l'objet du chapitre VIII qui présente des structures en chaîne. Une digression est faite avec les dispositifs à commutation de capacités, qui ne sont pas de type numérique au sens strict, mais qui sont néanmoins de type échantillonné et sont des compléments très utiles aux filtres numériques. Pour guider l'utilisateur, un résumé des mérites respectifs des structures décrites est donné en fin de chapitre.

Certains équipements, par exemple en instrumentation ou dans le domaine des télécommunications, font intervenir des signaux représentés par une suite de nombres complexes. Dans l'ensemble des signaux de ce type, une catégorie présente un intérêt pratique notable, celle des signaux analytiques. Leurs propriétés sont étudiées au chapitre IX, ainsi que la conception des dispositifs adaptés à la génération ou au traitement de tels signaux. Des notions complémentaires sur le filtrage sont également données dans ce chapitre, qui présente, d'une manière unifiée, les principales techniques d'interpolation.

Les machines de traitement numérique, quand elles fonctionnent en temps réel, opèrent à une cadence étroitement liée à la fréquence d'échantillonnage des signaux et leur complexité dépend du volume d'opérations à faire et de l'intervalle de temps disponible pour les réaliser. La fréquence d'échantillonnage des signaux est généralement imposée à l'entrée ou à la sortie des systèmes, mais à l'intérieur du système lui-même, il est possible de la faire varier pour l'adapter aux caractéristiques du signal et du traitement, et ainsi de réduire le volume d'opérations et la cadence des calculs. Une simplification des machines, qui peut être très importante, est obtenue en adaptant tout au long du traitement la fréquence d'échantillonnage à la largeur de bande du signal utile, c'est le filtrage multicadence présenté au chapitre X. Les incidences sur les caractéristiques du traitement sont décrites ainsi que les méthodes de réalisation. Des règles d'utilisation et d'évaluation sont fournies. Cette technique donne des résultats particulièrement intéressants pour les filtres à bande passante étroite ou pour la mise en œuvre d'ensembles appelés bancs de filtres. Dans ce dernier cas, le système associe à un ensemble de circuits déphaseurs un calculateur de Transformée de Fourier Discrète.

Les bancs de filtres pour la décomposition et la reconstruction des signaux sont devenus un outil de base pour la compression. Leur fonctionnement est décrit aux chapitres 11 et 12 avec les méthodes de calcul et les structures de réalisation.

Les filtres peuvent être déterminés à partir de spécifications dans le temps; c'est le cas par exemple de la modélisation d'un système comme décrit au chapitre 13. Si les caractéristiques varient, il peut être intéressant de modifier les coefficients en fonction des évolutions du système. Cette modification peut dépendre d'un critère d'approximation et se faire à une cadence qui peut atteindre la cadence d'échantillonnage du système; alors le filtre est dit adaptatif. Le chapitre 14 est consacré au filtrage adaptatif, dans le cas le plus simple, mais aussi le plus courant et le plus utile, celui où le critère d'approximation retenu est la minimisation de l'erreur quadratique moyenne et où les variations des coefficients se font suivant l'algorithme du gradient. Après un ensemble de rappels donnés au chapitre 13 sur les signaux aléatoires et leurs propriétés, en particulier la fonction et la matrice d'autocorrélation dont les valeurs propres jouent un rôle important, l'algorithme du gradient est présenté au chapitre 14 et ses conditions de convergence sont étudiées. Ensuite les deux paramètres d'adaptation principaux, la constante de temps et l'erreur résiduelle sont analysés, ainsi que la complexité arithmétique. Différentes structures de réalisation sont proposées.

Pour terminer, le chapitre 15 décrit brièvement quelques applications, en montrant comment les méthodes et techniques de base sont exploitées.

Chapitre 1

La numérisation du signal Échantillonnage et codage

La conversion d'un signal analogique sous forme numérique implique une double approximation. D'une part, dans l'espace des temps, le signal fonction du temps $s(t)$ est remplacé par ses valeurs $s(n\mathrm{T})$ à des instants multiples entiers d'une durée T; c'est l'opération d'échantillonnage. D'autre part, dans l'espace des amplitudes, chaque valeur $s(n\mathrm{T})$ est approchée par un multiple entier d'une quantité élémentaire q; c'est l'opération de quantification. La valeur approchée ainsi obtenue est ensuite associée à un nombre; c'est le codage, ce terme étant souvent utilisé pour désigner l'ensemble, c'est-à-dire le passage de la valeur $s(n\mathrm{T})$ au nombre qui la représente.

L'objet du présent chapitre est d'analyser l'incidence sur le signal de ces deux approximations.

Pour mener à bien cette tâche, on utilise deux outils de base qui sont l'analyse de Fourier et la théorie des distributions.

1.1 L'ANALYSE DE FOURIER

L'analyse de Fourier est un moyen de décomposer un signal en une somme de signaux élémentaires particuliers, qui ont la propriété d'êtres faciles à mettre en œuvre et à observer. L'intérêt de cette décomposition réside dans la fait que la réponse au signal d'un système obéissant au principe de superposition peut être déduite de la réponse aux signaux élémentaires. Ces signaux élémentaires sont périodiques et complexes, afin de permettre une étude en amplitude et en phase des systèmes; ils s'expriment par la fonction $s_e(t)$ telle que :

$$s_e(t) = e^{j2\pi ft} = \cos(2\pi ft) + j\sin(2\pi ft) \tag{1.1}$$

où f représente l'inverse de la période, c'est la fréquence du signal élémentaire.

Dans la mesure où les signaux élémentaires sont périodiques, il est clair que l'analyse se simplifie dans le cas où le signal est lui-même périodique. Ce cas va être examiné d'abord, bien qu'il ne corresponde pas aux signaux les plus intéressants, puisqu'un signal périodique est parfaitement déterminé et ne porte pratiquement pas d'information.

1.1.1 Développement en série de Fourier d'une fonction périodique

Soit $s(t)$, une fonction de la variable t périodique et de période T, c'est-à-dire satisfaisant la relation :

$$s(t + \mathrm{T}) = s(t) \tag{1.2}$$

Sous certaines conditions, on démontre que cette fonction est développable en série de Fourier, c'est-à-dire que l'égalité suivante est vérifiée :

$$s(t) = \sum_{n=-\infty}^{\infty} \mathrm{C}_n e^{j2\pi nt/\mathrm{T}} \tag{1.3}$$

L'indice n est un entier et les C_n sont appelés les coefficients de Fourier; ils sont définis par l'expression :

$$\mathrm{C}_n = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} s(t)\, e^{-j2\pi nt/\mathrm{T}}\, dt \tag{1.4}$$

En fait les coefficients de Fourier minimisent l'écart quadratique entre la fonction $s(t)$ et le développement (1.3). En effet la valeur (1.4) est obtenue en dérivant par rapport au coefficient d'indice n l'expression :

$$\int_0^{\mathrm{T}} \left(s(t) - \sum_{m=-\infty}^{\infty} \mathrm{C}_m e^{j2\pi mt/\mathrm{T}} \right)^2 dt$$

et en annulant cette dérivée.

Exemple : développement en série de Fourier de la fonction $i_p(t)$ constituée par une suite d'impulsions, séparées par la durée T, de largeur τ et d'amplitude a, centrée sur l'origine des temps (fig. 1.1).

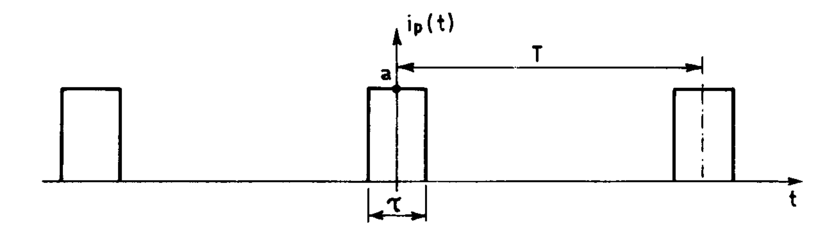


FIG. 1.1. *Suite d'impulsions*

Les coefficients C_n s'écrivent :

$$C_n = \frac{1}{T} \int_{-\tau/2}^{\tau/2} a e^{-j2\pi nt/T} \, dt = \frac{a\tau}{T} \; \frac{\sin\left(\pi n \frac{\tau}{T}\right)}{\pi n \frac{\tau}{T}} \tag{1.5}$$

et le développement est donné par :

$$i_p(t) = \frac{a\tau}{T} \sum_{n=-\infty}^{\infty} \frac{\sin\left(\pi n \frac{\tau}{T}\right)}{\pi n \frac{\tau}{T}} \; e^{j2\pi nt/T} \tag{1.6}$$

On imagine l'importance que prend cet exemple dans l'étude des systèmes échantillonnés.
Les propriétés des développements en série de Fourier sont présentées dans l'ouvrage [1]. Une propriété importante est exprimée par l'égalité de Bessel-Parseval qui traduit le fait que dans la décomposition du signal il y a conservation de la puissance :

$$\sum_{n=-\infty}^{\infty} |C_n|^2 = \frac{1}{T} \int_0^T |s(t)|^2 \, dt \tag{1.7}$$

Les signaux élémentaires qui résultent de la décomposition d'un signal périodique ont des fréquences qui sont des multiples entiers de $\frac{1}{T}$, l'inverse de la période ; ils couvrent un ensemble discret de l'espace des fréquences. Par contre si le signal n'est pas périodique, les signaux élémentaires résultant de la décomposition couvrent un domaine continu de l'espace des fréquences.

1.1.2 Transformation de Fourier d'une fonction

Soit $s(t)$ une fonction de la variable t ; sous certaines conditions on démontre l'égalité suivante :

$$s(t) = \int_{-\infty}^{\infty} S(f) \, e^{j2\pi ft} \, df \tag{1.8}$$

avec

$$S(f) = \int_{-\infty}^{\infty} s(t) \, e^{-j2\pi ft} \, dt \tag{1.9}$$

La fonction $S(f)$ est la transformée de Fourier de $s(t)$. Plus communément $S(f)$ est appelé spectre du signal $s(t)$.

Exemple : soit à calculer la transformée de Fourier $I(f)$ d'une impulsion isolée $i(t)$ de largeur τ, d'amplitude a et centrée sur l'origine des temps (fig. 1.2)

$$I(f) = \int_{-\infty}^{\infty} i(t)\, e^{-j2\pi ft}\, dt = a \int_{-\tau/2}^{\tau/2} e^{-j2\pi ft}\, dt$$

$$I(f) = a\tau\, \frac{\sin(\pi f\tau)}{\pi f\tau} \tag{1.10}$$

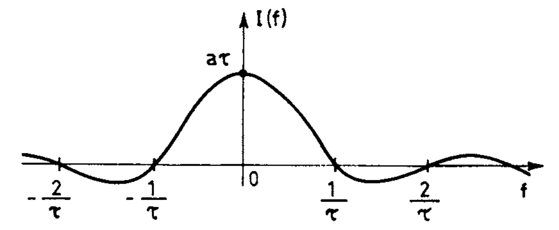


FIG. 1.2. *Impulsion isolée*

FIG. 1.3. *Spectre de l'impulsion isolée*

La figure 1.3 représente la fonction $I(f)$, qui sera très fréquemment utilisée par la suite. Il est important de remarquer qu'elle s'annule aux fréquences multiples entiers non nuls de l'inverse de la durée de l'impulsion.

L'Annexe 1 donne une tabulation de cette fonction.

La correspondance entre coefficients de Fourier et spectre apparaît nettement sur cet exemple. En effet, en rapprochant les relations (1.6) et (1.10) on vérifie que, au facteur $\frac{1}{T}$ près, les coefficients du développement en série de Fourier de la suite d'impulsions correspondent aux valeurs que prend le spectre de l'impulsion isolée aux fréquences multiples entiers de l'inverse de la période des impulsions.

En fait, on a la relation :

$$C_n = \frac{1}{T} S\left(\frac{n}{T}\right)$$

Une relation comparable à l'égalité de Bessel-Parseval existe pour une fonction non périodique. Dans ce cas, c'est non plus la puissance mais l'énergie du signal qui se trouve conservée :

$$\int_{-\infty}^{\infty} |S(f)|^2\, df = \int_{-\infty}^{\infty} |s(t)|^2\, dt \tag{1.11}$$

Soit $s'(t)$ la dérivée de la fonction $s(t)$; sa transformée de Fourier $S_d(f)$ s'écrit :

$$S_d(f) = \int_{-\infty}^{\infty} e^{-j2\pi ft} . s'(t)\, dt = j2\pi f . S(f) \tag{1.12}$$

Ainsi prendre la dérivée d'un signal amène une multiplication de son spectre par $j2\pi f$.

Une propriété essentielle de la transformation de Fourier, qui est en fait la principale raison de son utilisation, est qu'elle transforme un produit de convolu-

tion en un produit simple. En effet soit deux fonctions du temps $x(t)$ et $h(t)$ dont les transformées de Fourier sont respectivement $X(f)$ et $H(f)$. Le produit de convolution $y(t)$ est défini par :

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} x(t-\tau)\, h(\tau)\, d\tau \qquad (1.13)$$

La transformée de Fourier de ce produit s'écrit :

$$Y(f) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x(t-\tau)\, h(\tau)\, d\tau \right) e^{-j2\pi ft}\, dt$$

$$Y(f) = \int_{-\infty}^{\infty} h(\tau)\, e^{-j2\pi f\tau}\, d\tau . \int_{-\infty}^{\infty} x(u)\, e^{-j2\pi fu}\, du = H(f) . X(f)$$

Réciproquement, on montre que la transformée de Fourier d'un produit simple est un produit de convolution.

Un résultat intéressant pour l'étude de l'échantillonnage et se rapportant à l'exemple ci-dessus peut être déduit directement de ces propriétés. En effet soit à calculer la transformée de Fourier $II(f)$ de la fonction $i^2(t)$; d'après les relations (1.10) et (1.13), il vient :

$$II(f) = I(f) * I(f) = a . I(f) \qquad (1.14)$$

et par suite :

$$\int_{-\infty}^{\infty} \frac{\sin(\pi\varphi\tau)}{\pi\varphi\tau} . \frac{\sin[\pi(f-\varphi)\tau]}{\pi(f-\varphi)\tau}\, d\varphi = \frac{1}{\tau}\, \frac{\sin(\pi f\tau)}{\pi f\tau}$$

En prenant $f = \frac{n}{\tau}$, pour tout entier n non nul, on a :

$$\int_{-\infty}^{\infty} \frac{\sin(\pi\varphi\tau)}{\pi\varphi\tau} . \frac{\sin[\pi(\varphi\tau - n)]}{\pi(\varphi\tau - n)}\, d\varphi = 0 \qquad (1.15)$$

Les fonctions $\frac{\sin \pi(x-n)}{\pi(x-n)}$, avec n entier, forment un ensemble de fonctions orthogonales.

La définition et les propriétés de la transformation de Fourier s'étendent aux fonctions de plusieurs variables. Soit $s(x_1, x_2, \ldots, x_n)$ une fonction de n variables réelles, la transformée de Fourier est une fonction $S(\lambda_1, \lambda_2, \ldots, \lambda_n)$ définie par :

$$S(\lambda_1, \lambda_2, \ldots, \lambda_n) = \iint_{R^n} \cdots \int s(x_1, x_2, \ldots, x_n) e^{-j2\pi(\lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_n x_n)}\, dx_1\, dx_2, \ldots, dx_n \qquad (1.16)$$

Si la fonction $s(x_1, x_2, \ldots, x_n)$ est séparable c'est-à-dire si :
$s(x_1, x_2, \ldots, x_n) = s(x_1)\, s(x_2) \ldots s(x_n)$ alors il vient :

$$S(\lambda_1, \lambda_2, \ldots, \lambda_n) = S(\lambda_1)\, S(\lambda_2) \ldots S(\lambda_n)$$

Les variables $x_i (1 \leqslant i \leqslant n)$ représentent souvent des distances, par exemple dans le cas bidimensionnel, et les λ_i sont alors appelées fréquences spatiales.

Dans l'étude des signaux échantillonnés, la transformation de Fourier va être appliquée aux distributions.

1.2 LES DISTRIBUTIONS

Les distributions mathématiques constituent une définition mathématique correcte des distributions rencontrées en physique [1].

1.2.1 Définition

On appelle distribution D une fonctionnelle linéaire continue sur l'espace vectoriel $\mathcal{D}$ des fonctions définies sur $\mathbb{R}^n$, indéfiniment dérivables et à support borné.

A toute fonction φ appartenant à $\mathcal{D}$, la distribution D associe un nombre complexe $D(\varphi)$, qui sera aussi noté par $\langle D, \varphi \rangle$, avec les propriétés :

– $D(\varphi_1 + \varphi_2) = D(\varphi_1) + D(\varphi_2)$.
– $D(\lambda\varphi) = \lambda D(\varphi)$ où λ est un scalaire.
– Si φ_j converge vers φ quand j tend vers l'infini, la suite $D(\varphi_j)$ converge vers $D(\varphi)$.

Exemples :

• Si $f(t)$ est une fonction sommable sur tout ensemble borné, elle définit une distribution D_f par :

$$\langle D_f, \varphi \rangle = \int_{-\infty}^{\infty} f(t)\, \varphi(t)\, dt \tag{1.17}$$

• Si φ' désigne la dérivée de φ, la fonctionnelle :

$$\langle D, \varphi \rangle = \int_{-\infty}^{\infty} f(t)\, \varphi'(t)\, dt = \langle f, \varphi' \rangle \tag{1.18}$$

est une distribution.

• La distribution de Dirac δ est définie par :

$$\langle \delta, \varphi \rangle = \varphi(0) \tag{1.19}$$

La distribution de Dirac au point réel x est définie par :

$$\langle \delta(t - x), \varphi \rangle = \varphi(x) \tag{1.20}$$

On dit que cette distribution représente la masse + 1 au point x.

• Soit l'impulsion $i(t)$ de durée τ, d'amplitude $a = 1/\tau$, centrée sur l'origine. Elle définit une distribution D_i :

$$\langle D_i, \varphi \rangle = \frac{1}{\tau} \int_{-\tau/2}^{\tau/2} \varphi(t)\, dt$$

Pour des valeurs de τ très petites on obtient :

$$\langle D_i, \varphi \rangle \simeq \varphi(0)$$

c'est-à-dire que la distribution de Dirac peut être considérée comme la limite, quand τ tend vers 0, de la distribution D_i.

1.2.2 Dérivation des distributions

On définit la dérivée $\dfrac{\partial D}{\partial t}$ d'une distribution D par la relation :

$$\langle \frac{\partial D}{\partial t}, \varphi \rangle = -\langle D, \frac{\partial \varphi}{\partial t} \rangle \tag{1.21}$$

Soit par exemple la fonction Y de Heaviside, ou échelon unité, égale à 0 si $t < 0$ et $+1$ si $t \geqslant 0$.

$$\langle \frac{\partial Y}{\partial t}, \varphi \rangle = -\langle Y, \frac{\partial \varphi}{\partial t} \rangle = -\int_0^{\infty} \varphi'(t)\, \mathrm{d}t = \varphi(0) = \langle \delta, \varphi \rangle \tag{1.22}$$

Il en résulte que la discontinuité de Y apparaît sous la forme d'une masse ponctuelle unitaire dans sa dérivée.
Cet exemple illustre un intérêt pratique considérable de la notion de distribution, qui permet d'étendre aux fonctions discontinues un certain nombre de concepts et de propriétés des fonctions continues.

1.2.3 Transformation de Fourier d'une distribution

Par définition la transformée de Fourier d'une distribution D est une distribution notée FD telle que :

$$\langle FD, \varphi \rangle = \langle D, F\varphi \rangle \tag{1.23}$$

Par application de cette définition aux distributions à support ponctuel il vient :

$$\langle F\delta, \varphi \rangle = \langle \delta, F\varphi \rangle = \int_{-\infty}^{\infty} \varphi(t)\, dt = \langle 1, \varphi \rangle \tag{1.24}$$

Par suite : $F\delta = 1$.
De même $F\delta(t-a) = e^{-j2\pi fa}$.

Un cas fondamental pour l'étude de l'échantillonnage est celui que constitue la suite des distributions de Dirac décalées de T, notée u et telle que :

$$u(t) = \sum_{n=-\infty}^{\infty} \delta(t - n\mathrm{T}) \tag{1.25}$$

Cette suite est une distribution de masses unitaires aux points dont l'abscisse est un multiple entier de T. Sa transformée de Fourier s'écrit :

$$\mathrm{F}u = \sum_{n=-\infty}^{\infty} e^{-j2\pi fn\mathrm{T}} = \mathrm{U}(f) \tag{1.26}$$

On démontre que cette somme est en fait une distribution ponctuelle.

Une démonstration intuitive peut être obtenue à partir du développement en série de Fourier de la fonction $i_p(t)$ constituée par la suite d'impulsions séparées par la durée T, de largeur τ, d'amplitude $1/\tau$, centrée sur l'origine des temps.

En effet on peut considérer que : $u(t) = \lim\limits_{\tau \to 0} i_p(t)$.

En se reportant à la relation (1.6) on trouve : $\lim\limits_{\tau \to 0} i_p(t) = \dfrac{1}{\mathrm{T}} \sum\limits_{n=-\infty}^{\infty} e^{-j2\pi nt/\mathrm{T}}$

Il en résulte que : $$\mathrm{U}(f) = \sum_{n=-\infty}^{\infty} e^{-j2\pi fn\mathrm{T}} = \frac{1}{\mathrm{T}} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{\mathrm{T}}\right) \tag{1.27}$$

Cette propriété fondamentale démontrée dans l'ouvrage [1], ainsi que dans l'ouvrage [2], s'exprime comme suit :

La transformée de Fourier de la distribution temporelle comportant une masse unitaire en chaque point dont l'abscisse est un multiple entier de T est une distribution fréquentielle comportant la masse 1/T aux points dont l'abscisse est un multiple entier de 1/T.

Ce résultat va être utilisé pour étudier l'échantillonnage d'un signal.

La propriété que possède la transformation de Fourier d'échanger convolution et multiplication s'applique également aux distributions.

Avant d'étudier les incidences sur le signal des opérations d'échantillonnage et quantification, il est utile de caractériser les signaux qui sont les plus fréquemment traités.

1.3 LES PRINCIPAUX SIGNAUX TRAITÉS

Les signaux sont définis par une fonction du temps $s(t)$. Cette fonction peut être une expression analytique ou la solution d'une équation différentielle, auquel cas le signal est appelé déterministe.

1.3.1 Les signaux déterministes

Les signaux de ce type les plus utilisés sont les signaux sinusoïdaux ; par exemple :

$$s(t) = \mathrm{A} \cos(\omega t + \alpha)$$

où A est l'amplitude, $\omega = 2\pi f$ la pulsation et α la phase du signal.

Ils sont faciles à reproduire, à reconnaître aux différents points d'un système et offrent une possibilité de visualisation simple des caractéristiques. De plus, comme indiqué aux paragraphes précédents, ils servent de base à la décomposition d'un signal déterministe quelconque, par l'intermédiaire de la Transformation de Fourier.

Si le système considéré est linéaire et invariant dans le temps, il peut être caractérisé par sa réponse en fréquence $\mathrm{H}(\omega)$. Pour chaque valeur de la fréquence, $\mathrm{H}(\omega)$ est un nombre complexe dont le module est l'amplitude de la réponse. Par convention on désigne par phase de la réponse du système la fonction $\varphi(\omega)$ telle que :

$$\mathrm{H}(\omega) = |\mathrm{H}(\omega)|\, e^{-j\varphi(\omega)} \tag{1.28}$$

Cette convention permet d'exprimer le temps de propagation de groupe $\tau(\omega)$, fonction positive dans les systèmes réels, par :

$$\tau(\omega) = \frac{d\varphi}{d\omega} \tag{1.29}$$

Le temps de propagation de groupe fait référence aux lignes de transmission, sur lesquelles les différentes fréquences d'un signal se propagent à des vitesses différentes, ce qui entraîne une dispersion dans le temps de l'énergie du signal. Pour illustrer cette notion, soit deux fréquences proches $\omega \pm \Delta\omega$ auxquelles correspondent les phases par unité de longueur $\varphi \pm \Delta\varphi$. Le signal somme s'écrit :

$$s(t) = \cos[(\omega + \Delta\omega)t - (\varphi + \Delta\varphi)] + \cos[(\omega - \Delta\omega)t - (\varphi - \Delta\varphi)]$$

ou encore

$$s(t) = 2\cos(\omega t - \varphi)\cos(\Delta\omega t - \Delta\varphi)$$

C'est un signal modulé et il n'y a pas de dispersion si les deux facteurs subissent le même retard par unité de longueur, c'est-à-dire si $\Delta\varphi/\Delta\omega$ est une constante. Le temps de propagation de groupe caractérise donc la dispersion apportée à un signal par une ligne de transmission ou un système équivalent.

En appliquant au système le signal sinusoïdal $s(t)$, on obtient en sortie le signal résultant $s_r(t)$ tel que :

$$s_r(t) = \mathrm{A}\,.\,|\mathrm{H}(\omega)| \cos[\omega t + \alpha - \varphi(\omega)] \tag{1.30}$$

C'est encore un signal sinusoïdal et la comparaison avec le signal appliqué permet une visualisation de la réponse du système. On imagine aisément l'importance de cette procédure pour les opérations de test par exemple.

Les signaux déterministes cependant ne représentent pas très bien les signaux réels, car, en fait, ils ne portent pas d'information, si ce n'est pas leur présence

même. Les signaux réels sont généralement caractérisés par une fonction $s(t)$ aléatoire. Pour le test et l'analyse des systèmes on utilise aussi des signaux aléatoires, mais qui présentent des caractéristiques particulières pour ne pas compliquer exagérément la génération et l'exploitation. Une étude des signaux aléatoires est faite dans le tome 2 de la référence [2].

1.3.2 Les signaux aléatoires

Un signal aléatoire est défini à chaque instant t par la loi de probabilité de son amplitude $s(t)$. Cette loi peut s'exprimer par une densité de probabilité $p(x, t)$ définie comme suit :

$$p(x, t) = \lim_{\Delta x \to 0} \frac{\text{Proba}\,[x \leqslant s(t) \leqslant x + \Delta x]}{\Delta x} \qquad (1.31)$$

Il est stationnaire si ces propriétés statistiques sont indépendantes du temps, c'est-à-dire que sa densité de probabilité est indépendante du temps :

$$p(x, t) = p(x)$$

Il est du second ordre s'il possède un moment d'ordre 1 appelé valeur moyenne, qui est l'espérance mathématique de $s(t)$, notée $\mathrm{E}[s(t)]$ et définie par :

$$m_1(t) = \mathrm{E}[s(t)] = \int_{-\infty}^{\infty} x\,.\,p(x, t)\; dx$$

et un moment d'ordre 2, appelé fonction covariance :

$$\mathrm{E}[s(t_1)\,.\,s(t_2)] = m_2(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1\,.\,x_2\,.\,p(x_1, x_2; t_1, t_2)\; dx_1\, dx_2$$

où $p(x_1, x_2; t_1, t_2)$ est la densité de probabilité du couple de variables aléatoires $[s(t_1), s(t_2)]$.

Le caractère de stationnarité peut être limité aux moments du premier et du second ordre ; on dit alors que le signal est stationnaire d'ordre 2 ou stationnaire au sens large, et pour un tel signal il vient :

$$\mathrm{E}(s(t)] = \int_{-\infty}^{\infty} x\,.\,p(x)\; dx = m_1$$

L'indépendance du temps se traduit comme suit pour le densité de probabilité $p(x_1, x_2; t_1, t_2)$:

$$p(x_1, x_2; t_1, t_2) = p(x_1, x_2; 0, t_2 - t_1) = p(x_1, x_2; \tau)$$

avec

$$\tau = t_2 - t_1$$

Seul intervient l'écart entre les deux instants d'observation du signal :

$$\mathrm{E}[(s(t_1)\,.\,s(t_2)] = m_2(\tau) \qquad (1.33)$$

La fonction $r_{xx}(\tau)$ telle que :

$$r_{xx}(\tau) = \mathrm{E}[s(t).s(t-\tau)] \tag{1.34}$$

prend le nom de fonction d'autocorrélation du signal aléatoire, qu'elle caractérise.

Un signal aléatoire $s(t)$ possède aussi une moyenne temporelle m_T, qui est une variable aléatoire définie par :

$$m_\mathrm{T} = \lim_{\mathrm{T}\to\infty} \frac{1}{\mathrm{T}} \int_{-\mathrm{T}/2}^{\mathrm{T}/2} s(t)\, dt \tag{1.35}$$

L'ergodicité de cette moyenne exprime le fait qu'elle prend une valeur déterminée k avec la probabilité 1. Pour un signal stationnaire, l'ergodicité de la moyenne temporelle entraîne l'égalité avec la moyenne des amplitudes à un instant donné. En effet prenons l'espérance de la variable m_T :

$$\mathrm{E}[m_\mathrm{T}] = k = \mathrm{E}\left[\lim_{\mathrm{T}\to\infty} \frac{1}{\mathrm{T}} \int_{-\mathrm{T}/2}^{\mathrm{T}/2} s(t)\, dt\right] = \lim_{\mathrm{T}\to\infty} \frac{1}{\mathrm{T}} \int_{-\mathrm{T}/2}^{\mathrm{T}/2} \mathrm{E}[s(t)]\, dt = m_1$$

Ce résultat a des conséquences pratiques importantes puisqu'il fournit un moyen d'accéder aux propriétés statistiques du signal à un instant donné à partir de l'observation de ce signal au cours du temps.

L'ergodicité de la covariance dans le cas stationnaire est également très intéressante car elle conduit à la relation :

$$r_{xx}(\tau) = \lim_{\mathrm{T}\to\infty} \frac{1}{\mathrm{T}} \int_{-\mathrm{T}/2}^{\mathrm{T}/2} s(t)\, s(t-\tau)\, dt \tag{1.36}$$

La fonction d'autocorrélation du signal $s(t)$, $r_{xx}(\tau)$ est fondamentale pour l'étude des signaux stationnaires d'ordre deux ergodiques. Ses principales propriétés sont les suivantes :

– C'est une fonction paire :

$$r_{xx}(\tau) = r_{xx}(-\tau)$$

– Son maximum est à l'origine et correspond à la puissance du signal P :

$$r_{xx}(0) = \mathrm{E}[s^2(t)] = \mathrm{P}$$

– La densité spectrale de puissance est la tranformée de Fourier de la fonction d'autocorrélation :

$$\Phi_{xx}(f) = \int_{-\infty}^{\infty} r_{xx}(\tau)\, e^{-j2\pi f\tau}\, d\tau = 2\int_{0}^{\infty} r_{xx}(\tau)\cos(2\pi f\tau)\, d\tau$$

En effet : $r_{xx}(\tau) = s(\tau) * s(-\tau)$ et, si $\mathrm{S}(f)$ désigne la transformée de Fourier de $s(t)$, il vient :

$$\Phi_{xx}(f) = \mathrm{S}(f).\overline{\mathrm{S}(f)} = |\mathrm{S}(f)|^2 \tag{1.37}$$

Cette dernière propriété se traduit physiquement par le fait que plus le signal est à variation rapide, c'est-à-dire plus son spectre s'étend vers les fréquences éle-

vées, plus sa fonction d'autocorrélation est étroite. A la limite le signal est purement aléatoire et la fonction s'annule pour $\tau \neq 0$. On se trouve en présence d'un signal appelé bruit blanc, et tel que :

$$r_{xx}(\tau) = \mathrm{P}\delta$$

Alors la densité spectrale est constante :

$$\Phi_{xx}(f) = \mathrm{P}$$

En fait un tel signal n'a pas de réalité physique puisque sa puissance est infinie, mais il constitue un modèle mathématique commode pour les signaux dont la densité spectrale est quasi constante sur une large bande de fréquence.

1.3.3 Les signaux gaussiens

Parmi les lois de probabilité que l'on peut considérer pour un signal $s(t)$, il est une catégorie qui présente un grand intérêt, celle des lois normales ou lois de Gauss. En effet les distributions aléatoires normales conservent leur caractère normal dans toute opération linéaire, par exemple la convolution par une distribution certaine, le filtrage, la dérivation ou l'intégration. Aussi ces distributions aléatoires sont-elles très utilisées pour la modélisation des signaux réels et le test des systèmes.

Une variable aléatoire x est dite gaussienne si sa loi de probabilité a une densité $p(x)$ qui suit la loi normale ou loi de Gauss :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-m)^2}{2\sigma^2}} \qquad (1.38)$$

La valeur m est la moyenne de la variable x ; la variance σ^2 est le moment d'ordre deux de la variable centrée $(x - m)$; σ est aussi appelé l'écart-type.

La variable $\left(\dfrac{x-m}{\sigma}\right)$ est dite réduite, elle a une moyenne nulle et un écart-type unité. Une tabulation et une représentation très utile sous forme de courbe sont fournies en annexe II.

Une variable aléatoire est caractérisée par la loi de probabilité de son amplitude, mais aussi par l'ensemble de ses moments m_n, tels que :

$$m_n = \int_{-\infty}^{\infty} x^n\, p(x)\, dx \qquad (1.39)$$

Ces moments sont les coefficients du développement en série entière d'une fonction $\mathrm{F}(u)$ appelée fonction caractéristique de la variable aléatoire x et définie par :

$$\mathrm{F}(u) = \int_{-\infty}^{\infty} e^{jux}\, p(x)\, dx \qquad (1.40)$$

C'est, à un changement de variable près, la transformée de Fourier inverse de la densité de probabilité $p(x)$ et l'on a également :

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-jux}\, \mathrm{F}(u)\, du \tag{1.41}$$

À partir de la relation (1.40) on obtient le développement en série entière suivant :

$$\mathrm{F}(u) = \sum_{n=0}^{\infty} \frac{(ju)^n}{n!}\, m_n \tag{1.42}$$

Et pour une variable gaussienne centrée :

$$\mathrm{F}(u) = e^{-\frac{1}{2}\sigma^2 u^2} \tag{1.43}$$

Par développement en série et identification avec (1.42), on obtient tous les moments :

$$m_{2n} = \frac{(2n)!}{n!\,2^n}\, \sigma^{2n}$$

Par exemple, pour $n = 2$, on obtient $m_4 = 3\sigma^4$. Tous les moments d'ordre impair d'une variable gaussienne centrée sont nuls, d'après la définition de la loi de probabilité elle-même.

La loi normale se généralise aux variables aléatoires à plusieurs dimensions [3]. La fonction caractéristique d'une variable gaussienne à k dimensions $x(x_1, \ldots, x_k)$ s'écrit :

$$\mathrm{F}(u_1, \ldots, u_k) = e^{-\frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} r_{ij} u_i u_j} \tag{1.44}$$

avec :

$$r_{ij} = \mathrm{E}(x_i x_j)$$

La densité de probabilité est obtenue par transformation de Fourier. Dans le cas à 2 dimensions il vient :

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - r^2}}\, e^{-\frac{1}{2(1-r^2)}\left[\frac{x_1^2}{\sigma_1^2} - \frac{2r x_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right]} \tag{1.45}$$

où r désigne le coefficient de corrélation :

$$r = \frac{\mathrm{E}(x_1 x_2)}{\sigma_1\sigma_2}$$

Un signal aléatoire $s(t)$ est dit gaussien, si pour un ensemble de k instants $t_i (1 \leqslant i \leqslant k)$ la variable aléatoire à k dimensions $s = [s(t_1), \ldots, s(t_k)]$ est gaussienne.

D'après la relation (1.44), la loi de probabilité de cette variable est complètement définie par la fonction d'auto-corrélation $r_{xx}(\tau)$ du signal $s(t)$.

Exemple :

Le signal défini par les équations suivantes :

$$r_{xx}(\tau) = \sigma^2 e^{-\frac{|\tau|}{RC}} \tag{1.46}$$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}} \tag{1.47}$$

est une approximation d'un bruit gaussien blanc d'utilisation courante dans l'analyse des systèmes ou la modélisation des signaux. C'est un signal stationnaire de moyenne nulle dont la densité spectrale n'est pas rigoureusement constante, mais correspond à une répartition uniforme filtrée par un filtre passe-bas de type RC. Il s'obtient par amplification du bruit d'agitation thermique aux bornes d'une résistance.

La distribution normale peut être obtenue à partir d'une distribution de probabilité $p(x)$ uniforme sur l'intervalle [0, 1]. En effet soit $p(y)$ la distribution dite de Rayleigh :

$$p(y) = \frac{y}{\sigma^2}\, e^{-\frac{y^2}{2\sigma^2}}\ ; \quad y \geqslant 0 \tag{1.48}$$

qui a pour moment d'ordre deux ou puissance, $2\sigma^2$, pour moyenne $\sqrt{\frac{\pi}{2}}\sigma$ et pour variance $\left(2 - \frac{\pi}{2}\right)\sigma^2$. Par un changement de variable tel que :

$$p(x)\,dx = p(y)\,dy$$

il vient :

$$p(y) = p(x)\,\frac{dx}{dy} = \frac{dx}{dy}\ ; \quad \left|\frac{dx}{dy}\right| = \frac{y}{\sigma^2}\, e^{-\frac{y^2}{2\sigma^2}}$$

d'où :

$$x = e^{-\frac{y^2}{2\sigma^2}}\ ; \quad y = \sigma\sqrt{2\ln\left(\frac{1}{x}\right)} \tag{1.49}$$

La distribution normale est obtenue en considérant deux variables y et x indépendantes et en posant :

$$z = y\cos 2\pi x \tag{1.50}$$

La démonstration fait intervenir la variable :

$$z' = y\sin 2\pi x$$

En effet, en utilisant la correspondance entre coordonnées polaires et cartésiennes, on peut écrire :

$$p(z, z')\,dz\,dz' = p(z)\,p(z')\,dz\,dz' = p(y)\,p(x)\,dx\,dy = p(z)\,p(z')\,ydy\,2\pi\,dx$$

d'où :

$$p(z)\,p(z') = \frac{1}{2\pi}\;\frac{1}{\sigma^2}\;e^{-\frac{y^2}{2\sigma^2}} = \frac{1}{2\pi\sigma^2}\;e^{-\frac{z^2+z'^2}{2\sigma^2}}$$

et finalement :

$$p(z) = \frac{1}{\sigma\sqrt{2\pi}}\;e^{-\frac{z^2}{2\sigma^2}}$$

Cette procédure est couramment utilisée pour produire des signaux Gaussiens numériques.

1.3.4 Facteur de crête d'un signal aléatoire

Un signal aléatoire est défini à chaque instant par une loi de probabilité de son amplitude, souvent telle que cette amplitude n'est pas bornée. C'est le cas des signaux gaussiens, comme le montre la relation (1.38).

Or le traitement d'un signal ne peut se réaliser que pour une gamme d'amplitudes limitée et des opérations de cadrage interviennent. Un paramètre important est le facteur de crête défini pour le signal comme le rapport d'une certaine amplitude A_m à la valeur efficace σ. Par convention cette amplitude A_m est souvent prise comme la valeur qui n'est pas dépassée pendant plus de 10^{-5} du temps. Ce rapport est exprimé en décibels (dB) par F_c tel que :

$$F_c = 20\log\left[\frac{A_m}{\sigma}\right] \tag{1.51}$$

où log désigne le logarithme en base 10.

Pour un signal gaussien le facteur de crête est de 12,9 dB. Appliquée à un signal sinusoïdal cette définition conduit à un facteur de crête de 3 dB.

Un modèle stationnaire utilisé pour représenter le signal téléphonique est constitué par le signal aléatoire dont la densité de probabilité des amplitudes suit la loi exponentielle, ou de Laplace, suivante :

$$p(x) = \frac{1}{\sigma\sqrt{2}}\;e^{-\sqrt{2}\,\frac{|x|}{\sigma}} \tag{1.52}$$

Le facteur de crête dans ce cas s'élève à 17,8 dB.

En conclusion, les fonctions aléatoires stationnaires d'ordre deux ergodiques, caractérisées par une loi de probabilité des amplitudes et une fonction d'autocorrélation, permettent de modéliser la plupart des signaux à traiter et sont très utilisées dans l'étude et l'analyse des systèmes.

En plus des possibilités de représentation des signaux il est important de pouvoir disposer d'une mesure globale, par exemple afin de pouvoir suivre un signal au cours du traitement. Une telle mesure est obtenue en définissant des normes sur la fonction qui représente le signal.

1.4 NORMES D'UNE FONCTION

Une norme est une fonction positive réelle, qui vérifie les relations :

$$\|x\| \geqslant 0; \quad k\|x\| = \|kx\|$$

où k est un réel positif.

Une catégorie très utilisée de normes est l'ensemble des normes dites normes-Lp [4] :

La norme-Lp d'une fonction continue $s(t)$ définie sur l'intervalle [0, 1] est notée $\|s\|_p$ et définie par :

$$\|s\|_p = \left[\int_0^1 |s(t)|^p \, dt\right]^{\frac{1}{p}} \tag{1.53}$$

Trois valeurs de p sont intéressantes :

– $p = 1$:

$$\|s\|_1 = \int_0^1 |s(t)| \, dt \tag{1.53-a}$$

– $p = 2$:

$$\|s\|_2^2 = \int_0^1 |s(t)|^2 \, dt \tag{1.53-b}$$

c'est l'expression de l'énergie du signal $s(t)$

– $p = \infty$:

$$\|s\|_\infty = \max_{0 \leqslant t \leqslant 1} |s(t)| \tag{1.53-c}$$

Cette norme est aussi appelée norme de Tchebycheff. Les normes sont utilisées également dans les techniques d'approximation pour mesurer l'écart entre une fonction $f(x)$ et la fonction à approcher $F(x)$. L'approximation est faite au sens des moindres carrés si la norme L_2 est utilisée et au sens de Tchebycheff si la norme L_∞ est utilisée.

Les normes -Lp peuvent être généralisées par l'introduction d'une fonction de pondération réelle positive $p(x)$. La norme-Lp pondérée de la fonction d'écart $f(x) - F(x)$ s'écrit alors :

$$\|f(x) - F(x)\|_p = \left[\int_0^1 |f(x) - F(x)|^p \, p(x) \, dx\right]^{\frac{1}{p}} \tag{1.53-d}$$

Ces notions sont appliquées dans le calcul des coefficients des filtres et aussi des facteurs d'échelle qui commandent les cadrages des données dans les mémoires.

1.5 L'OPÉRATION D'ÉCHANTILLONNAGE

L'échantillonnage consiste à représenter un signal fonction du temps $s(t)$ par ses valeurs $s(nT)$ à des instants multiples entiers d'une durée T, appelée période d'échantillonnage. Une telle opération s'analyse de façon simple et concise par l'intermédiaire de la théorie des distributions. En effet, par définition, la distribution de masses unitaires aux points de l'axe réel multiples entiers de la période T, associe à la fonction $s(t)$ l'ensemble de ses valeurs $s(nT)$ où n est un entier. Conformément aux notations précédemment retenues cette distribution est notée $u(t)$ et s'écrit :

$$u(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT)$$

L'opération d'échantillonnage affecte le spectre $S(f)$ du signal. Considérant la relation fondamentale (1.27), il apparaît que le spectre $U(f)$ de la distribution $u(t)$ est constitué de raies d'amplitude $\frac{1}{T}$ aux fréquences qui sont des multiples entiers de la fréquence d'échantillonnage $fe = \frac{1}{T}$. Par suite $u(t)$ s'exprime comme une somme de signaux élémentaires :

$$u(t) = \frac{1}{T} \sum_{n=-\infty}^{\infty} e^{j2\pi nt/T} \tag{1.54}$$

Alors la suite des valeurs de signal $s(nT)$ correspond au produit de l'ensemble des signaux élémentaires qui constituent $u(t)$ par le signal $s(t)$. C'est-à-dire que physiquement, l'opération d'échantillonnage est une modulation en amplitude par le signal d'une infinité de porteurs à des fréquences qui sont des multiples entiers de la fréquence d'échantillonnage $f_e = 1/T$. Par suite le spectre du signal échantillonné comprend la fonction $S(f)$, désignée par la bande de base, ainsi que les bandes images qui correspondent à la translation de la bande de base de multiples entiers de la fréquence d'échantillonnage.

L'opération d'échantillonnage et son incidence sur le spectre du signal sont représentées sur la figure 1.4.

FIG. 1.4. *Incidence spectrale de l'échantillonnage*

Le spectre du signal échantillonné $S_e(f)$ a pour expression le produit de convolution de $S(f)$ par $U(f)$ soit :

$$S_e(f) = \frac{1}{T} \sum_{n=-\infty}^{\infty} S\left(f - \frac{n}{T}\right) \tag{1.55}$$

Il est important de remarquer que la fonction $S_e(f)$ est périodique, c'est-à-dire que l'échantillonnage a introduit une périodicité dans l'espace des fréquences, ce qui constitue une caractéristique fondamentale des signaux échantillonnés.

L'opération d'échantillonnage telle qu'elle vient d'être décrite et que l'on désigne par échantillonnage idéal, peut sembler peu réaliste, dans la mesure où il apparaît difficile dans la réalité d'atteindre, de manipuler ou de restituer une valeur d'un signal à un instant ponctuel; les échantillonneurs réels ou les circuits qui restituent les échantillons possèdent un certain temps d'ouverture. En fait on peut montrer que l'échantillonnage ou la restitution d'échantillons par des impulsions ayant une largeur donnée, introduit simplement une modification du spectre du signal.

En effet dans l'opération d'échantillonnage du signal $s(t)$ par la suite d'impulsions séparées par la durée T, de largeur τ et d'amplitude a, il se peut que l'on recueille à la période n une quantité σ_n qui s'écrit :

$$\sigma_n = a \int_{nT-\tau/2}^{nT+\tau/2} s(t)\, dt$$

Cette quantité exprime le résultat de la convolution du signal $s(t)$ par l'impulsion élémentaire $i(t)$ et la fonction dont on prélève dans ce cas les valeurs aux instants d'échantillonnage nT est la fonction $s * i$; c'est-à-dire que le signal échantillonné a pour spectre non pas $S(f)$ mais le produit :

$$S(f)\,.\,a\tau\,.\,\frac{\sin(\pi f \tau)}{\pi f \tau}$$

Le raisonnement est le même pour le cas de la restitution d'échantillons avec une durée τ. En fait c'est le produit de convolution des échantillons $s(nT)$ avec l'impulsion élémentaire $i(t)$ qui est restitué.

D'où la proposition :

L'échantillonnage ou la restitution d'échantillons par des impulsions de largeur τ peut être traité comme un échantillonnage idéal ou une restitution idéale, à la condition de multiplier le spectre du signal par le spectre de l'impulsion élémentaire.

En pratique dès que τ est faible devant la période T la correction devient négligeable.

1.6 L'ÉCHANTILLONNAGE EN FRÉQUENCE

L'échantillonnage considéré ci-dessus est de type temporel. Cependant les propriétés énoncées sont aussi applicables à un échantillonnage de type fréquentiel.

Calculons le spectre d'une fonction périodique $s_p(t)$ de période T. Une telle fonction peut être considérée comme résultant du produit de convolution de la fonction $s(t)$, qui prend les valeurs de $s_p(t)$ sur une période et s'annule en dehors, et de la distribution ponctuelle $u(t)$. Il en résulte la relation suivante entre les transformées de Fourier :

$$S_p(f) = U(f) \,.\, S(f) = \frac{1}{T} \sum_{n=-\infty}^{\infty} S\left(\frac{n}{T}\right) \delta\left(f - \frac{n}{T}\right) \qquad (1.56)$$

En fait on retrouve les coefficients du développement en série de Fourier de la fonction $s_p(t)$. Le cas où $s(t)$ est une impulsion est représenté sur la figure 1.5.

Il apparaît que le spectre de la fonction périodique $s_p(t)$ est un spectre de raies qui constituent un échantillonnage du spectre de la fonction prise sur une période. L'échantillonnage dans l'espace des fréquences correspond à une périodicité dans l'espace des temps. Cette interprétation est utile dans l'analyse numérique des spectres.

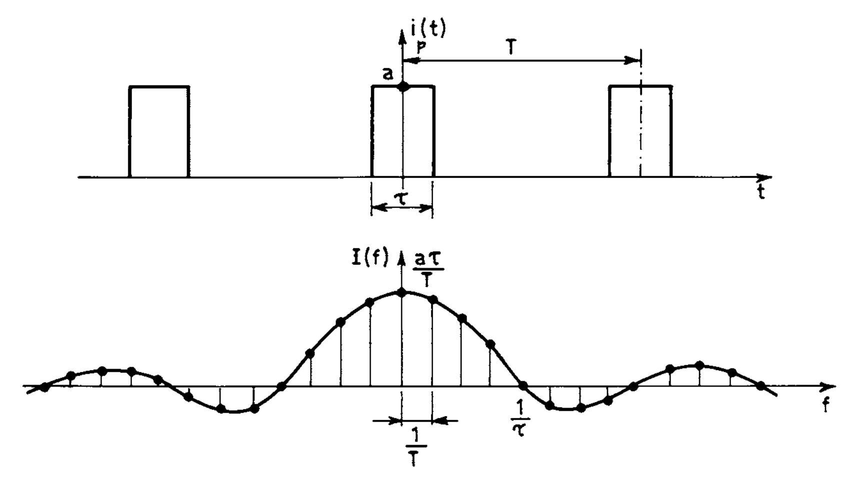


FIG. 1.5. *Spectre d'une suite d'impulsions*

1.7 LE THÉORÈME DE L'ÉCHANTILLONNAGE

Ce théorème exprime les conditions dans lesquelles la suite des échantillons d'un signal représente correctement ce signal. Un signal est supposé être correctement représenté par la suite de ses échantillons prélevés avec la périodicité T s'il est possible, à partir de cette suite de valeurs, de restituer intégralement le signal d'origine.

L'échantillonnage a introduit une périodicité du spectre dans l'espace des fréquences ; restituer le signal d'origine, c'est supprimer cette périodicité, c'est-à-dire

éliminer les bandes images, opération qui peut être réalisée à l'aide d'un filtre passe bas dont la fonction de transfert $H(f)$ vaut $1/f_e$ jusqu'à la fréquence $f_e/2$ et 0 aux fréquences supérieures. En sortie d'un tel filtre apparaît un signal continu, qu'il est possible d'exprimer en fonction des valeurs $s(nT)$. La réponse impulsionnelle du filtre $h(t)$ s'écrit, d'après la relation (1.10) :

$$h(t) = \frac{\sin(\pi t/T)}{\pi t/T}$$

Le signal de sortie du filtre, $s(t)$, correspond au produit de convolution de la suite $s(nT)$ par la fonction $h(t)$, soit :

$$s(t) = \int_{-\infty}^{\infty} \left[\sum_{n=-\infty}^{\infty} s(\theta)\, \delta(\theta - nT) \right] \frac{\sin \pi(t-\theta)/T}{\pi(t-\theta)/T}\, d\theta$$

d'où :

$$s(t) = \sum_{n=-\infty}^{\infty} s(nT) \frac{\sin \pi(t/T - n)}{\pi(t/T - n)} \tag{1.57}$$

C'est la formule de calcul des valeurs du signal aux instants situés entre les échantillons. Pour les multiples de la période T elle fournit bien $s(nT)$. Le processus de reconstitution du signal est représentée sur la figure 1.6.

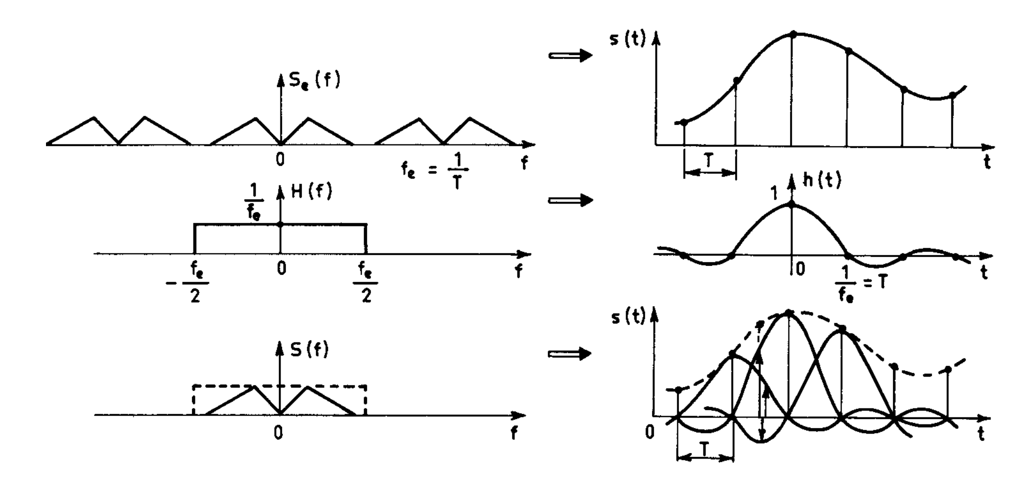


FIG. 1.6. *Reconstitution du signal après échantillonnage*

Pour que le signal calculé $s(t)$ soit identique au signal d'origine, il faut que le spectre $S(f)$ soit identique au spectre du signal d'origine. Comme le montre la figure 1.6 cette condition est vérifiée si et seulement si le spectre d'origine ne contient pas de composantes aux fréquences supérieures ou égales à $f_e/2$.

Si ce n'est pas le cas, les bandes images chevauchent la bande de base comme sur la figure 1.7, on dit qu'il y a repliement de bande, et le filtre de restitution four-

nit un signal différent du signal d'origine. D'où le théorème de l'échantillonnage ou théorème de Shannon :

Un signal qui ne contient pas de composantes à des fréquences supérieures ou égales à une valeur *fm* est entièrement déterminé par la suite de ses valeurs à des instants régulièrement espacés de la durée T = $\mathbf{\frac{1}{2fm}}$.

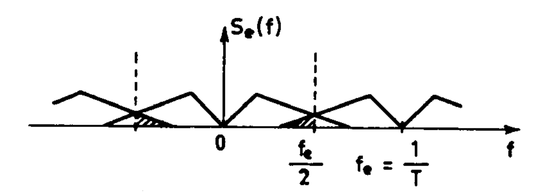


FIG. 1.7. *Repliement de bande*

La fréquence d'échantillonnage d'un signal est ainsi déterminée par la limite supérieure de sa bande de fréquence. Dans la pratique on limite généralement par filtrage la bande du signal avant échantillonnage à la fréquence f_e, à une valeur inférieure à $fe/2$ pour que le filtre de restitution soit réalisable.

Il est intéressant de remarquer que la fréquence d'échantillonnage d'un signal est en fait déterminée par la largeur de bande qu'il occupe ; en effet la restitution a été illustrée sur la figure 1.6 pour un signal basse fréquence auquel a été associé un filtre passe-bas. On conçoit que le même raisonnement s'applique aussi à un signal occupant un domaine limité de l'espace des fréquences auquel serait associé un filtre passe-bande. Cette propriété est applicable en particulier aux signaux modulés et est utilisée dans certains types de filtres numériques.

Le résultat donné à la fin du paragraphe 1.1.2 permet de présenter l'échantillonnage sous un autre aspect. En effet, la relation (1.57) montre que l'échantillonnage correspond à une décomposition du signal $s(t)$ suivant l'ensemble des fonctions orthogonales $\frac{\sin \pi(t/T - n)}{\pi(t/T - n)}$ et le théorème de Shannon exprime simplement la condition pour que cet ensemble forme une base de décomposition du signal.

1.8 ÉCHANTILLONNAGE DE SIGNAUX SINUSOÏDAUX ET DE SIGNAUX ALÉATOIRES

Les propriétés énoncées ci-dessus sont bien illustrées par l'échantillonnage de signaux sinusoïdaux, dont les particularités sont utilisables dans de nombreuses applications.

1.8.1 Signaux sinusoïdaux

Soit le signal $s(t) = \cos(2\pi ft + \varphi)$, avec $0 \leqslant \varphi \leqslant \dfrac{\pi}{2}$, échantillonné avec la période $T = 1/f_e = 1$.

Les échantillons sont donnés par la suite $s(n)$ telle que :

$$s(n) = \cos(2\pi fn + \varphi)$$

Si le rapport $f/f_e = f$ est un nombre rationnel, il vient :

$$f = N1/N2 \quad \text{avec} \quad N1 \quad \text{et} \quad N2 \text{ entiers.}$$

Alors :

$$s(n + N2) = \cos[2\pi f(n + N2) + \varphi] = s(n)$$

La suite $s(n)$ présente la périodicité N2 et comprend au plus N2 nombres différents. D'autre part la fréquence d'échantillonnage étant supérieure au double de la fréquence du signal, on a nécessairement : $N1/N2 < \dfrac{1}{2}$. L'ensemble de N2 échantillons différents permet de représenter un nombre de signaux sinusoïdaux égal au plus grand entier inférieur à N2/2. Par exemple si N2 = 8, avec l'ensemble des nombres : $2\cos\left(2\pi \dfrac{n}{8} + \varphi\right)$, $(n = 0, 1, \ldots, 7)$, il est possible de représenter les échantillons des 3 signaux sinusoïdaux :

$$2\cos\left(2\pi \frac{N1}{8} t + \varphi\right) \quad \text{avec} \quad N1 = 1, 2, 3$$

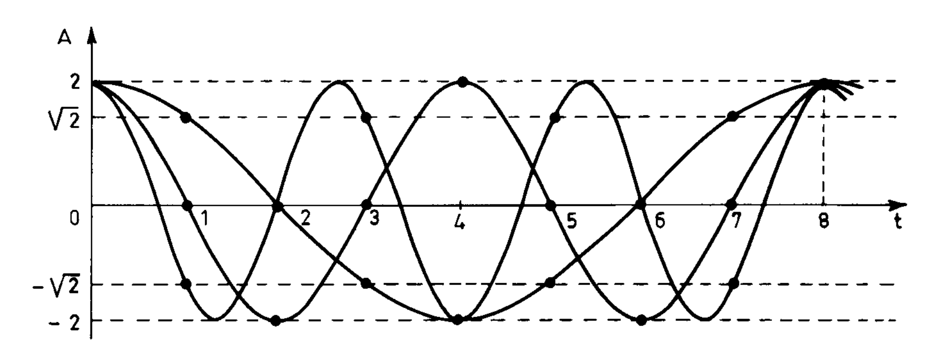


FIG. 1.8. *Échantillonnage des signaux :* $\cos\left(2\pi \dfrac{N}{8} t\right)$

La figure 1.8 représente cet échantillonnage pour $\varphi = 0$; dans ce cas il suffit même de 4 nombres : ± 2 et $\pm \sqrt{2}$.

Si l'on ajoute aux trois signaux sinusoïdaux de la figure 1.8, le signal continu de valeur 1 et le signal à la fréquence 1/2 d'amplitude 1 qui s'écrit : cos (πt), l'échantillonnage de cette somme donne des valeurs nulles, sauf aux instants multiples de 8, où la valeur 8 est obtenue, comme le montre la figure 1.9.*a*. Le spectre de cette somme est obtenu directement en appliquant la relation :

$$\cos x = \frac{1}{2}\,(e^{jx} + e^{-jx}).$$

Il est formé de raies d'amplitude 1 aux fréquences multiples de 1/8 (fig. 1.9.*b*). Or ce spectre a été étudié au paragraphe 1.2, et l'on peut constater que la relation (1.27) se trouve vérifiée.

La possibilité d'engendrer une gamme de signaux sinusoïdaux à partir d'un ensemble limité de nombres, stockés par exemple dans une mémoire, est utilisée dans les synthétiseurs de fréquence numériques.

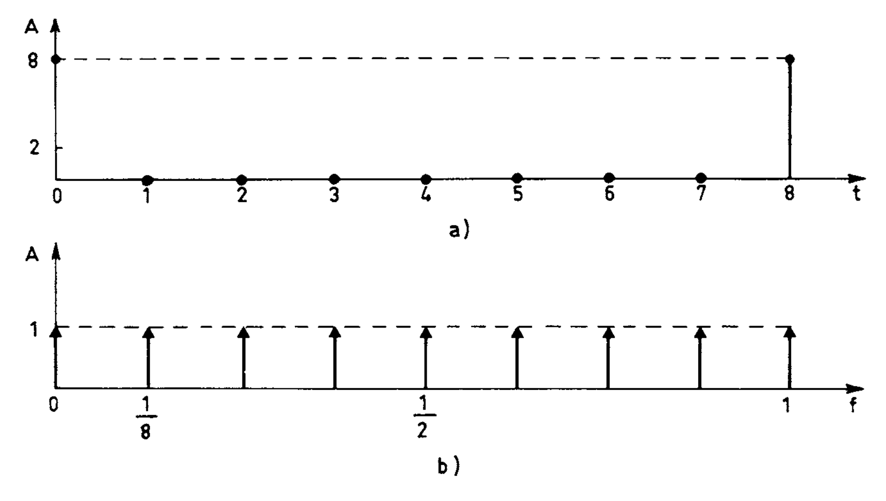


FIG. 1.9. *a*) *Échantillonnage du signal* $s(t) = 1 + 2\sum_{n=1}^{3} \cos\left(2\pi \frac{n}{8} t\right) + \cos(\pi t)$

b) *Spectre correspondant*

1.8.2 Signaux aléatoires discrets

Si le signal aléatoire $s(t)$ est échantillonné avec la période supposée unitaire T = 1, il en résulte un signal aléatoire discret $s(n)$, qui a par définition la même loi de probabilité de l'amplitude. Les résultats obtenus dans le cas continu se transposent au cas discret, en particulier pour les signaux aléatoires stationnaires du second ordre ergodiques [5].

Ainsi la fonction d'auto-corrélation du signal discret $s(n)$ est la suite $r(n)$ telle que :

$$r(n) = \mathrm{E}[s(i).s(i-n)] \tag{1.58}$$

C'est un échantillonnage de la fonction d'auto-corrélation $r_{xx}(\tau)$ du signal aléatoire continu définie par l'expression (1.34). Sa transformée de Fourier donne la densité spectrale énergétique $\Phi_e(f)$ du signal discret, qui est liée à la densité spectrale $\Phi_{xx}(f)$ du signal continu par une relation analogue à (1.55), c'est-à-dire :

$$\Phi_e(f) = \frac{1}{\mathrm{T}} \sum_{n=-\infty}^{\infty} \Phi_{xx}\left(f - \frac{n}{\mathrm{T}}\right) \tag{1.59}$$

Si la fréquence d'échantillonnage n'a pas une valeur suffisante, ou si le spectre $\Phi_{xx}(f)$ s'étend sur un domaine non borné, il y a repliement.

L'hypothèse d'ergodicité pour le signal discret $s(n)$ conduit à la relation :

$$r(n) = \lim_{\mathrm{N}\to\infty} \frac{1}{2\mathrm{N}+1} \sum_{i=-\mathrm{N}}^{\mathrm{N}} s(i)\, s(i-n) \tag{1.60}$$

Cette relation permet d'appliquer la notion de fonction d'auto-corrélation aux signaux déterministes. Ainsi, pour un signal périodique et de période N_0, la fonction d'auto-corrélation est la suite $r(n)$ définie par :

$$r(n) = \frac{1}{\mathrm{N}_0} \sum_{i=0}^{\mathrm{N}_0-1} s(i)\, s(i-n) \tag{1.61}$$

C'est une suite périodique et de même période.

Exemple :

$$s(n) = \mathrm{A} \sin\left(2\pi \frac{n}{\mathrm{N}_0}\right)$$

$$r(n) = \frac{1}{\mathrm{N}_0} \sum_{i=0}^{\mathrm{N}_0-1} \mathrm{A}^2 \sin\left(2\pi \frac{i}{\mathrm{N}_0}\right) \sin\left(2\pi \frac{i-n}{\mathrm{N}_0}\right)$$

$$r(n) = \frac{\mathrm{A}^2}{2} \cos\left(2\pi \frac{n}{\mathrm{N}_0}\right)$$

On retrouve bien la puissance du signal pour $r(0)$ et la périodicité N_0.

Un signal aléatoire discret peut aussi être défini en tant que tel. Par exemple si la suite $r(n)$ s'annule pour $n \neq 0$, le signal $s(n)$ est un bruit blanc discret dont la densité spectrale est constante sur l'intervalle de fréquence $\left[-\frac{1}{2}, \frac{1}{2}\right]$. Ce signal a une réalité physique, c'est une suite de variables aléatoires non corrélées ; pour l'obtenir il suffit de faire appel à un algorithme qui fournit des nombres statistiquement indépendants.

1.8.3 Génération d'un bruit discret

La génération de nombres aléatoires figure généralement au catalogue des fonctions des calculateurs scientifiques. Il est ainsi possible en logiciel de former une suite de nombres, utilisable comme signal de test en traitement numérique.

Au paragraphe 1.8.1 on a montré qu'il est particulièrement simple de produire numériquement des signaux sinusoïdaux; de tels signaux peuvent aussi servir à simuler un bruit, par exemple par addition d'un grand nombre de sinusoïdes de fréquences différentes, d'amplitude constante et de phase aléatoire ou pseudo-aléatoire. Cette approche peut conduire à des réalisations particulièrement simples, comme la méthode qui a été normalisée pour l'appareillage de mesure utilisé dans les transmissions téléphoniques numériques. Cette méthode consiste à engendrer une séquence pseudo-aléatoire, qui est une suite périodique de $2^N - 1$ bits comprenant à une unité près autant de «zéros» que de «uns» et qui simule une suite aléatoire dans laquelle les bits seraient indépendants et auraient la probabilité 1/2 de valoir «zéro» ou «un», ou pour centrer les variables, de valoir – 1/2 ou + 1/2.

Si une opération de filtrage, qui en fait consiste en une sommation pondérée, est effectuée sur une telle suite, les nombres obtenus après filtrage suivent une loi de probabilité qui s'approche de la loi normale.

Les séquences pseudo-aléatoires sont étudiées dans la référence [6], elles sont facilement obtenues à l'aide d'un registre à décalage à N bits, convenablement bouclé. La figure 1.10 donne un exemple, utilisé en appareillage de mesure, où N = 17. Le polynôme générateur s'écrit :

$$g(x) = 1 + x^3 + x^{17} \tag{1.62}$$

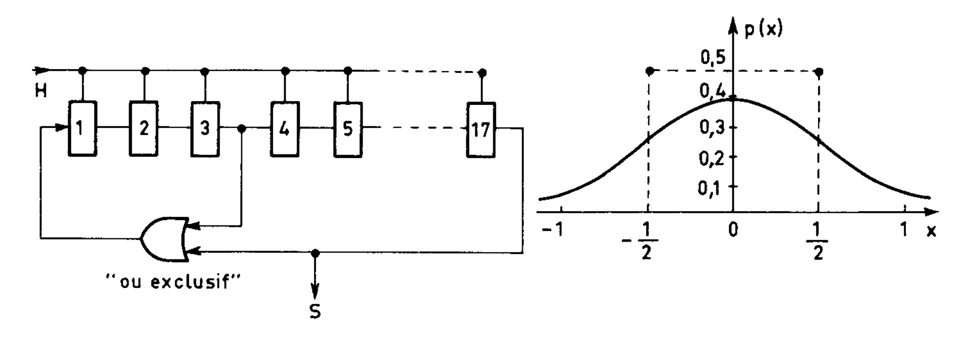


FIG. 1.10. *Générateur de séquence pseudo-aléatoire et loi de probabilité après filtrage*

la suite comprend $2^N - 1 = 131\,071$ bits, elle est périodique et de période $T = (2^N - 1) . \tau$, si $\tau = \dfrac{1}{f_H}$ désigne la période de l'horloge du circuit. Le spectre est formé de raies distantes de $\dfrac{1}{T}$. Pour $f_H = 370$ kHz, l'espacement entre deux raies est de 2,8 Hz et l'on trouve 36 raies dans 100 Hz.

En opérant sur cette suite un filtrage à bande étroite qui ne conserve que la bande 450-550 Hz on obtient un signal approchant les caractéristiques gaussiennes, dont le facteur de crête est de 10,5 dB et qui constitue un excellent signal de test pour les équipements de transmission numérique. Si le filtrage est fait numériquement la suite de nombres obtenue peut être utilisée pour tester les équipements de traitement numérique.

1.9 L'OPÉRATION DE QUANTIFICATION

La quantification est l'approximation de chaque valeur du signal s (t) par un multiple entier d'une quantité élémentaire q, appelée échelon de quantification. Si q est constant quelle que soit l'amplitude du signal, la quantification est dite uniforme. Cette opération revient à faire passer le signal dans un organe qui possède une caractéristique en marche d'escalier, comme le montre la figure 1.11 pour q = 1, et fournit le signal sq (t).

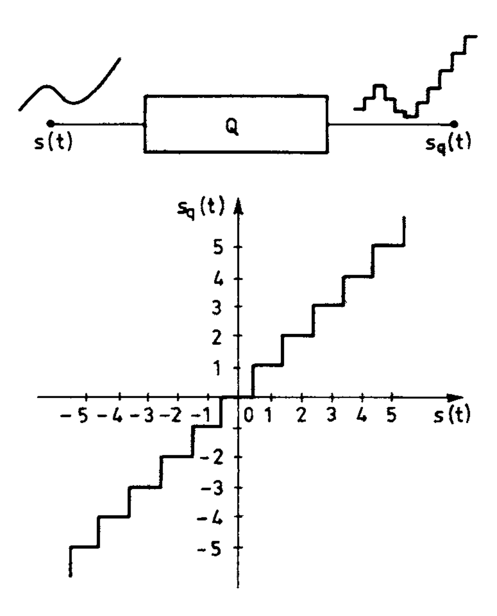


FIG. 1.11. *L'opération de quantification*

La manière dont l'approximation est faite définit le centrage de cette caractéristique. Par exemple la figure représente le cas, appelé arrondi, où toute valeur du signal comprise entre $(n - 1/2)q$ et $(n + 1/2)q$ est arrondie à nq. C'est l'approximation par défaut, qui est désignée, quand elle porte sur des nombres, par troncation et qui consiste à approcher par nq toute la valeur comprise entre nq et $(n + 1)q$; la caractéristique se déplace alors de $q/2$ vers la droite sur l'axe des abscisses.

L'effet de cette approximation est de superposer au signal d'origine un signal d'erreur $e(t)$ désigné par distorsion de quantification ou plus communément par bruit de quantification ; il vient :

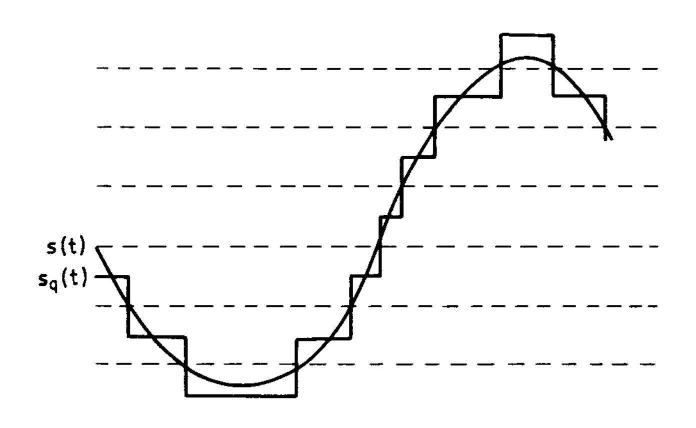
$$s(t) = s_q(t) + e(t) \tag{1.63}$$

Une illustration est donnée par la figure 1.12 dans le cas de l'arrondi. Les amplitudes multiples impairs de $q/2$ sont appelées amplitudes de décision.

L'amplitude du signal d'erreur est comprise entre $-\ q/2$ et $q/2$. Sa puissance mesure la dégradation que subit le signal.

Quand les variations du signal sont grandes par rapport à l'échelon de quantification, c'est-à-dire que la quantification est faite avec suffisamment de finesse, le signal d'erreur est équivalent à un ensemble de signaux élémentaires, constitués chacun par un segment de droite (fig. 1.13). La puissance d'un tel signal élémentaire de durée τ s'écrit :

$$\mathrm{B} = \frac{1}{\tau}\int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} e^2(t)\, dt = \frac{1}{\tau}\left(\frac{q}{\tau}\right)^2 \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} t^2\, dt = \frac{q^2}{12} \tag{1.64}$$



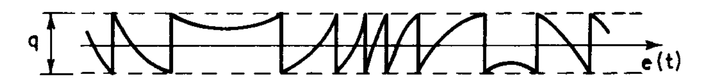
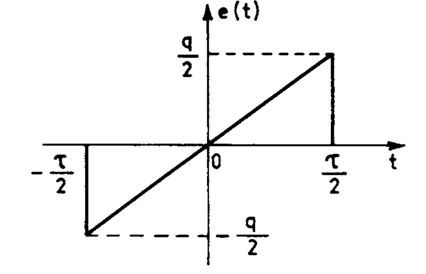


FIG. 1.12. *Erreur de quantification*



FIG. 1.13. *Signal d'erreur élémentaire*



La valeur ainsi obtenue, $B = \frac{q^2}{12}$, est une estimation de la puissance du bruit de quantification suffisante dans la plupart des cas réels.

La distribution spectrale du signal d'erreur est plus difficile à cerner. Le spectre du signal d'erreur élémentaire de la figure 1.13, $E_\tau(f)$, peut être calculé à partir de celui de sa dérivée. Ainsi en utilisant les relations (1.22) puis (1.12), on obtient :

$$E_\tau(f) = \frac{1}{j2\pi f} \cdot q \cdot \left[\frac{\sin(\pi f \tau)}{\pi f \tau} - \cos(\pi f \tau) \right] \qquad (1.65)$$

Il apparaît que la plus grande partie de l'énergie se trouve au voisinage de la fréquence $\frac{1}{\tau}$. Dans ces conditions la répartition spectrale du signal d'erreur dépend d'une part de la pente du signal élémentaire, c'est-à-dire en fait de la distribution statistique de la dérivée du signal $s'(t)$, et d'autre part de la grandeur de l'échelon de quantification q par rapport au signal. La référence [7] donne le calcul de ce spectre pour un signal de bruit et fait apparaître un étalement en fonction de la fréquence, quand le pas de quantification est suffisamment petit, sur un domaine qui couvre plusieurs centaines de fois la largeur de bande du signal. Si le signal à quantifier n'est pas un signal aléatoire, le spectre du signal d'erreur peut se concentrer sur certaines fréquences par exemple les harmoniques d'un signal sinusoïdal.

Dans la conversion d'un signal analogique sous forme numérique, la quantification intervient conjointement avec l'échantillonnage, ces deux opérations étant réalisées successivement. Bien que l'échantillonnage soit en général fait en premier, il est équivalent de faire la quantification d'abord et l'échantillonnage ensuite, à une fréquence f_e habituellement un peu supérieure au double de la largeur de bande du signal. Dans ces conditions le signal d'erreur a souvent un spectre qui s'étend bien au-delà de la fréquence d'échantillonnage, et comme c'est en réalité la somme du signal et du signal d'erreur qui est échantillonnée, le phénomène de repliement du spectre intervient et la totalité de l'énergie du signal d'erreur se retrouve dans la bande de fréquences $[-f_e/2, f_e/2]$. La plupart du temps les conditions sont remplies pour que la densité spectrale énergétique du bruit de quantification soit constante et l'on retiendra le résultat suivant :

Le bruit produit dans l'opération de quantification uniforme avec un échelon q a une puissance qui s'exprime en général par $B = q^2/12$ et présente une répartition spectrale constante dans la bande de fréquences $[-f_e/2, f_e/2]$.

Il faut remarquer que la quantification des petits signaux, ceux dont l'amplitude est de l'ordre de grandeur de l'échelon q, dépend beaucoup du centrage de la caractéristique. Par exemple avec le centrage de la figure 1.11 un signal sinusoïdal d'amplitude inférieure à $q/2$ est totalement supprimé. Il est possible cependant de coder convenablement ces petits signaux en leur superposant un signal auxiliaire de grande amplitude qui est éliminé par la suite.

Le codage d'un signal introduit ainsi une limitation pour les faibles amplitudes mais il impose également une borne aux fortes amplitudes.

1.10 LA DYNAMIQUE DE CODAGE

Le signal échantillonné et quantifié en amplitude est représenté par une suite de nombres presque toujours sous forme binaire. Si chaque nombre compte N bits, le nombre maximum d'amplitudes quantifiées qu'il est possible de distinguer s'élève à 2N. Alors la gamme des amplitudes qu'il est possible de coder est soumise à une double limitation : vers les faibles valeurs elle se trouve limitée par l'échelon de quantification q et vers les fortes valeurs par 2N. q. Toute amplitude qui dépasse cette valeur ne peut être représentée et il y a écrêtage du signal. Il s'en suit une dégradation, par exemple par distorsion harmonique si le signal est sinusoïdal.

Si la gamme des amplitudes à coder couvre le domaine $[-A_m, +A_m]$, il vient :

$$A_m = 2^N . q/2 \tag{1.66}$$

et d'autre part, avec l'arrondi, le signal d'erreur $e(t)$ est tel que :

$$|e(t)| \leqslant A_m . 2^{-N}$$

On appelle puissance de crête d'un codeur la puissance du signal sinusoïdal ayant l'amplitude maximale admissible sans écrêtage, A_m. Elle s'exprime par :

$$P_c = \frac{1}{2}\left[\frac{2^N . q}{2}\right]^2 = 2^{2N-3} . q^2$$

La figure 1.14 représente ce signal avec le pas de quantification et les amplitudes de décision.

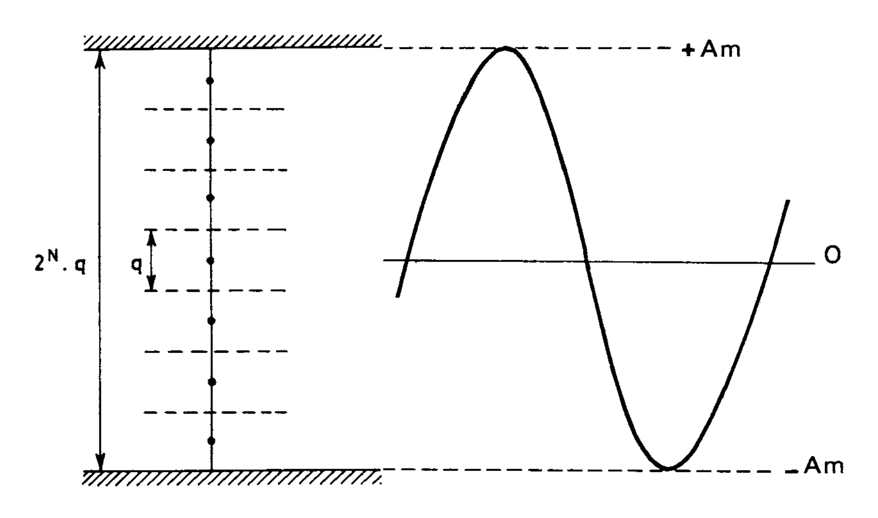


FIG. 1.14. *Puissance de crête du codeur*

On définit la dynamique du codage comme le rapport de cette puissance de crête à la puissance du bruit de quantification ; c'est en fait le rapport signal à bruit

maximal pour un signal sinusoïdal avec codage uniforme. Cette dynamique s'exprime par la formule suivante :

$$P_c/B = (S/B)_{max} = 2^{2N-3}.12 = 3/2.2^{2N}$$

ou plus commodément en décibels :

$$P_c/B = 6{,}02\ N + 1{,}76\ dB \qquad (1.67)$$

Cette formule, d'une grande utilité pratique, relie le nombre de bits du codage à la plage des amplitudes qui peuvent être codées.

Très souvent cependant le signal à coder n'est pas un signal sinusoïdal. Il est possible toutefois de se ramener à ce cas si, pour le signal à coder, est définie une puissance équivalente de crête, qui est alors prise comme puissance de crête du codeur. Le cas des signaux aléatoires gaussiens, est particulièrement important car ils représentent convenablement beaucoup de signaux rencontrés en pratique. Il faut alors positionner correctement l'amplitude maximale du codeur par rapport à l'amplitude du signal, de façon à ce que la distorsion introduite par écrêtage reste dans les limites imposées.

En examinant le tableau donné en annexe 2, on peut remarquer que la probabilité pour que l'amplitude d'un signal de moyenne nulle et de puissance σ^2 dépasse 3,4 σ est inférieure à 10^{-3}. La figure 1.15 donne un exemple de codage avec $\sigma = q$. Il apparaît que la probabilité d'écrêtage est inférieure à 5.10^{-4} avec les valeurs de paramètres choisies.

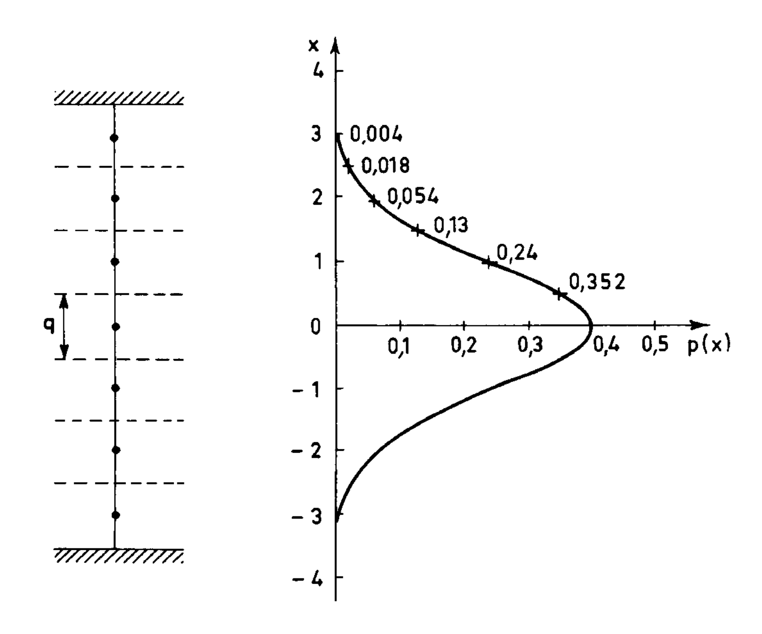


FIG. 1.15. *Codage d'un signal gaussien*

Finalement, pour obtenir le rapport signal à bruit maximal associé à un signal donné, il faut considérer le rapport signal à bruit de crête :

$$(S/B)_c = \frac{A_m^2 q^2}{12} = 3 \,.\, 2^N$$

et soustraire le facteur de crête F_c. L'expression générale du rapport signal à bruit maximal pour un signal quelconque est donc la suivante :

$$(S/B)_{max} = 6{,}02\ N + 4{,}77 - F_c\ dB \qquad (1.67bis)$$

Ce résultat est utilisé non seulement dans la spécification des codeurs de signaux analogiques mais aussi dans le traitement numérique pour la détermination des mémoires de données et le cadrage des nombres.

La dynamique de codage, pour un nombre de bits donné, peut être considérablement augmentée si le codage est fait avec un échelon de quantification qui varie avec l'amplitude du signal; c'est le codage non linéaire. De nombreuses lois de variation peuvent être envisagées. Cependant il en est une qui est particulièrement importante puisqu'elle a été normalisée par l'Union Internationale des Télécommunications (UIT) pour le codage de signaux téléphonique dans les réseaux de télécommunications, c'est la loi segmentée à 13 segments [8].

1.11 CODAGE NON LINÉAIRE SUIVANT UNE LOI SEGMENTÉE

Dans l'opération de codage non linéaire suivant la loi segmentée à 13 segments, les amplitudes positives et négatives à coder sont divisées en 7 plages, à chacune desquelles est associé un échelon de quantification dont la grandeur résulte de la multiplication d'un échelon élémentaire q par une puissance de 2. Cette opération peut être considérée comme résultant d'un codage linéaire précédé d'une compression selon laquelle le signal x est transformé en signal y conformément aux relations suivantes :

$$\begin{aligned} y &= \text{signe}\,(x)\,.\,\frac{1 + \ln A\,|x|}{1 + \ln A} \quad \text{pour} \quad \frac{1}{A} \leqslant |x| \leqslant 1 \\ y &= \text{signe}\,(x)\,.\,\frac{A\,|x|}{1 + \ln A} \quad \text{pour} \quad 0 \leqslant |x| \leqslant \frac{1}{A} \end{aligned} \qquad (1.68)$$

Le paramètre A détermine l'augmentation de la dynamique du codeur; la valeur retenue est A = 87,6. Finalement la caractéristique de compression suivant la loi A à 13 segments est donnée par la figure 1.16 et décrite comme suit :

si $0 \leqslant |x| \leqslant \frac{1}{64}$, alors : $y = 16x$

$\frac{1}{64} \leqslant |x| \leqslant \frac{1}{32}$ $\quad y = 8x + 1/8$

$$\frac{1}{32} \leqslant |x| \leqslant \frac{1}{16} \qquad y = 4x + 1/4$$

$$\frac{1}{16} \leqslant |x| \leqslant \frac{1}{8} \qquad y = 2x + 3/8$$

$$\frac{1}{8} \leqslant |x| \leqslant 1/4 \qquad y = x + 1/2$$

$$\frac{1}{4} \leqslant |x| \leqslant 1/2 \qquad y = \frac{1}{2}x + 5/8$$

$$\frac{1}{2} \leqslant |x| \leqslant 1 \qquad y = \frac{1}{4}x + 3/4$$

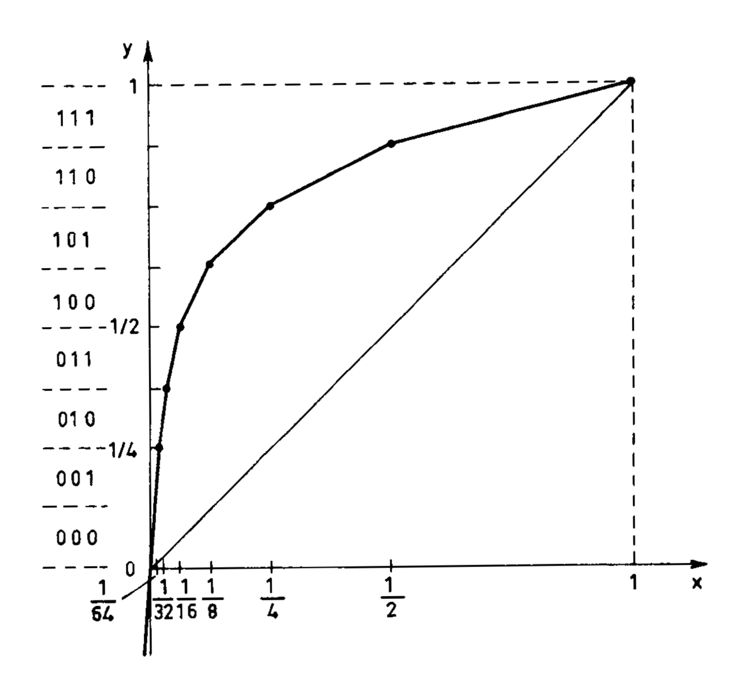


Fig. 1.16. *Caractéristique de compression à 13 segments*

Cette caractéristique fait apparaître 7 segments de droite dans le quadrant positif et dans le quadrant négatif ; les 2 segments qui entourent l'origine étant colinéaires, au total la caractéristique compte bien 13 segments.

La quantification des amplitudes y étant faite avec l'échelon q, celle des amplitudes x voisines de l'origine est faite avec l'échelon $q/16$, c'est-à-dire que la dynamique du codeur se trouve augmentée de 24 dB. Les amplitudes voisines de l'unité sont moins bien quantifiées puisque l'échelon se trouve multiplié par 4. La

puissance du bruit de quantification est fonction de l'amplitude du signal : pour chaque valeur il faut calculer un échelon moyen faisant intervenir la statistique du signal.

La figure 1.17 donne le rapport signal à bruit en fonction du niveau du signal après codage, pour le codage à 8 bits linéaire et non linéaire d'un signal gaussien.

Le niveau de référence pour le signal (0 dB) est la puissance de crête du codeur. On remarque l'extension de la dynamique due au codage non linéaire. Pour les amplitudes faibles la quantification correspond en fait à 12 bits. En réalité le signal codé suivant la loi non linéaire peut être obtenu à partir d'une quantification à 12 bits, suivie d'un traitement numérique qui est très proche de la conversion d'un nombre entier en un nombre à virgule flottante :

Par exemple au nombre à 12 bits : + 0 0 0 1 0 1 1 0 1 1 0
correspond le nombre à 8 bits : + 1 0 0 0 1 1 0

par application de la loi de compression.

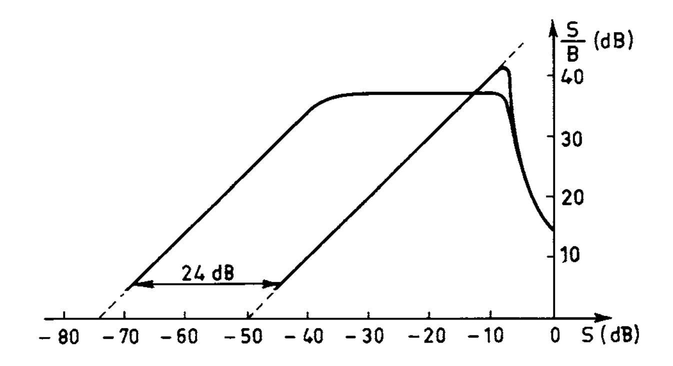


FIG. 1.17. *Codage à 8 bits linéaire et non linéaire d'un signal gaussien*

Les trois bits qui suivent le signe donnent le code du segment, ou exposant ; les quatre bits suivants indiquent la position sur le segment, ou mantisse. La différence avec la conversion entier-virgule flottante apparaît au voisinage de l'origine.

Ce traitement nécessite pour sa mise en œuvre soit un réseau de portes c'est la réalisation parallèle, soit un registre à décalage associé à un compteur à 3 bits dans la réalisation série. On peut également utiliser une mémoire où est stockée la table de conversion.

Une autre loi de codage non linéaire est également utilisée en télécommunications, dite loi μ à 15 segments. Elle correspond à la relation de compression suivante :

$$y = \text{signe}\,(x)\,.\,\frac{\ln(1+\mu|x|)}{\ln(1+\mu)} \quad \text{pour} \quad -1 \leqslant x \leqslant 1 \qquad (1.69)$$

La valeur retenue pour le paramètre de compression est $\mu = 255$.

1.12 OPTIMISATION DU CODAGE

En poursuivant dans la voie du perfectionnement du codage, si la densité de probabilité p (x) de l'amplitude du signal est connue, on peut déterminer la caractéristique de quantification qui, pour un nombre de bits N donné, minimise la puissance de la distorsion totale.

Dans l'opération de quantification, la plage des amplitudes du signal est divisée en $M = 2^N$ plages élémentaires (x_{i-1}, x_i) avec $-\frac{M}{2} + 1 \leqslant i \leqslant \frac{M}{2}$, et chaque plage élémentaire est représentée par une valeur y_i, comme le montre la figure 1.18. L'optimisation consiste à déterminer l'ensemble des valeurs x_i et y_i qui minimise la puissance du signal d'erreur E^2 donnée par :

$$E^2 = \sum_{i=-\frac{M}{2}+1}^{\frac{M}{2}} \int_{x_{i-1}}^{x_i} (x - y_i)^2 \, p(x) \, dx$$

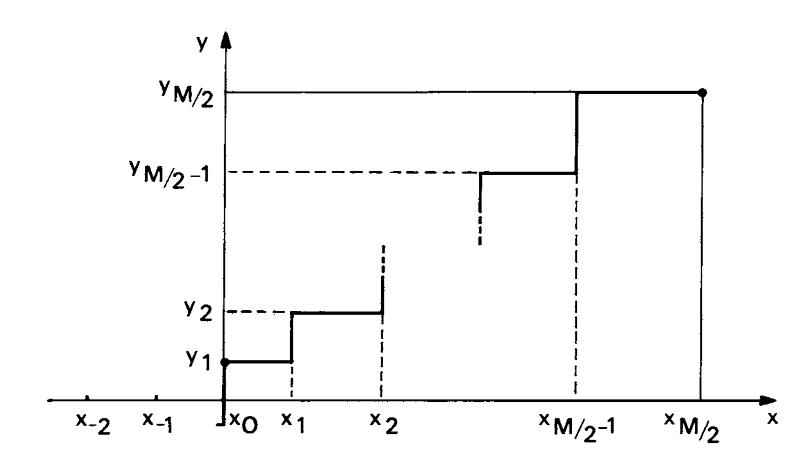


FIG. 1.18. *Caractéristique de quantification optimale*

En prenant la dérivée par rapport aux variables x_i et y_i, on montre que l'optimum est obtenu si les relations suivantes sont vérifiées :

$$x_i = \frac{1}{2}(y_i + y_{i+1}) \quad \text{pour} \quad -\frac{M}{2} + 1 \leqslant i \leqslant \frac{M}{2} - 1$$

$$\int_{x_{i-1}}^{x_i} (x - y_i) \, p(x) \, dx = 0 \quad \text{pour} \quad -\frac{M}{2} + 1 \leqslant i \leqslant \frac{M}{2} \qquad (1.70)$$

$$p\left(x_{\frac{M}{2}}\right) = p\left(x_{-\frac{M}{2}}\right) = 0$$

Ces relations permettent de déterminer la caractéristique de quantification. Si $p(x)$ est une fonction paire, on prend $x_0 = 0$ et l'on procède par itérations en choisissant a priori une valeur de y_1. Si la relation (1.70) n'est pas satisfaite pour $\frac{M}{2}$, on reprend les calculs pour une autre valeur de y_1 et ainsi de suite [9].

Le tableau 1.1 donne la puissance du signal d'erreur E^2 obtenue avec un signal gaussien de puissance unité, pour différentes valeurs du nombre de bits N, d'une part avec le codage optimal, d'autre part avec un codage à échelon constant pour le meilleur cadrage de la caractéristique de quantification [9]. Le tableau 1.2, tiré de la référence [10], donne, dans les mêmes conditions, les valeurs correspondant à un signal à densité de probabilité exponentielle selon la relation (1.48).

Tableau 1.1. – Codage d'un signal gaussien unitaire.

E^2 \ N	1	2	3	4	5
Codage optimal	0,3634	0,1175	0,03454	0,0095	0,0025
Codage à échelon constant	0,3634	0,1188	0,03744	0,01154	0,00349
Entropie H	1	1,911	2,825	3,765	4,730

Tableau 1.2. – Codage d'un signal laplacien unitaire.

E^2 \ N	1	2	3	4	5
Codage optimal	0,5	0,1765	0,0548	0,0154	0,00414
Codage à échelon constant	0,5	0,1963	0,0717	0,0254	0,0087

L'optimisation du codage peut aussi être abordée sous l'aspect du contenu informationnel en introduisant la fonction entropie H définie par [2, Tome 3] :

$$H = -\sum_i p_i \log 2\,(p_i) \tag{1.71}$$

avec $-\frac{M}{2} + 1 \leqslant i \leqslant \frac{M}{2}$ et où p_i désigne la probabilité pour que l'amplitude du signal se trouve dans la plage représentée par la valeur y_i.

Compte tenu de la relation :

$$\sum_i p_i = 1$$

l'entropie est nulle lorsque l'amplitude du signal se concentre sur une seule plage et elle est maximale lorsque cette amplitude est répartie uniformément ; dans ce cas elle prend la valeur H_{max} égale au nombre de bits N du codage :

$$H_{max} = \log 2\,(M) = N \tag{1.72}$$

où log 2 (M) représente le logarithme en base 2, ou binal, du nombre M.

En fait l'entropie mesure l'écart d'une distribution de probabilité avec la distribution uniforme.

La caractéristique de quantification qui maximise l'entropie est donc celle qui conduit à des plages élémentaires correspondant à une distribution de probabilité uniforme.

La dernière ligne du tableau 1.1 montre que pour un signal gaussien, le codage qui minimise la puissance du signal d'erreur conduit à des valeurs de l'entropie proches du maximum N.

1.13 QUANTITÉ D'INFORMATION ET CAPACITÉ D'UN CANAL

Les résultats obtenus sur l'échantillonnage et la quantification peuvent être utilisés, à l'inverse, pour évaluer la quantité d'information portée par un signal ou pour déterminer la capacité d'un canal de transmission.

Un canal réel de largeur de bande f_m peut transporter $2f_m$ échantillons indépendants par seconde, comme le montre la figure 1.3, en remplaçant τ par $2f_m$. Quant à la quantité d'information par échantillon, elle dépend des puissances respectives du signal utile et du bruit, et de leurs distributions d'amplitude. Un cas particulier important est celui du canal gaussien [11].

Soit à transmettre un ensemble de M symboles de N bits chacun dans un canal en présence d'un bruit blanc gaussien de puissance $\sigma_b^2 = B$.

Dans un hyperespace à M dimensions, les M symboles occupent le volume d'une hypersphère V_M, défini par

$$V_M = \int_0^R r^{M-1}\,dr \int_{i=1} \cdots \int_{M-1} f(\theta_i)\,d\theta_i = \frac{R^M}{M}\,F(\theta) \tag{1.73}$$

En supposant une répartition uniforme des symboles dans l'hypersphère de rayon R, le signal correspondant a pour énergie :

$$E_s = \frac{1}{V_M}\int_0^R r^2 r^{M-1}\,dr \int_{i=1} \cdots \int_{M-1} f(\theta_i)\,d\theta_i = R^2\,\frac{M}{M+2} \tag{1.74}$$

La quantité d'information transmise est de MN bits. À chaque ensemble de bits possible, on peut associer dans l'hypersphère un volume V_s égal à :

$$V_s = \frac{V_M}{2^{MN}} = \frac{1}{M} F(\theta)\left(\frac{R}{2^N}\right)^M \tag{1.75}$$

À chaque ensemble de bits est associé un bruit à M composantes dont l'énergie s'écrit : $E_b = M\sigma_b^2$. Quand M tend vers l'infini, le point représentatif du bruit dans l'hypersphère se rapproche d'une sphère de rayon $\sqrt{M}\sigma_b$ et centrée sur le point représentatif de l'ensemble des bits. En effet, pour M variables aléatoires gaussiennes, $b(n)$, la variable $r = \sqrt{\sum_{n=1}^{M} b^2(n)}$ a pour moment d'ordre 1 : $m_1 = \sigma_b\sqrt{\frac{M^2}{M+1}}$ et sa variance $m_2 - m_1^2$ tend vers zéro quand M tend vers l'infini. Cette sphère a pour volume :

$$V_b = \frac{(\sqrt{M}\sigma_b)^M}{M} F(\theta) \tag{1.76}$$

Pour que la transmission se fasse sans erreur, il faut que cette sphère soit incluse dans le volume V_s attribué à chaque ensemble de bits, c'est-à-dire :

$$\sqrt{M}\sigma_b < \frac{R}{2^N} \tag{1.75}$$

Or, quand M tend vers l'infini, d'après la relation (1.74), R^2 représente l'énergie de l'ensemble du signal, de puissance S et du bruit, soit

$$R^2 = M(S + \sigma_b^2) \tag{1.76}$$

La condition de transmision sans erreur s'écrit alors

$$2^{2N} < \frac{S + \sigma_b^2}{\sigma_b^2} \tag{1.77}$$

d'où :

$$N < \frac{1}{2} \log 2 \left(1 + \frac{S}{\sigma_b^2}\right) \tag{1.78}$$

Si le canal est réel et a une largeur de bande W et s'il est sans distorsion, les symboles peuvent être émis à la cadence 2W et la capacité asymptotique du canal s'écrit, en bits par seconde :

$$C = W \log 2 \left(1 + \frac{S}{B}\right) \tag{1.79}$$

Il faut bien noter qu'une telle capacité suppose :
– un canal sans distorsion,
– un bruit blanc gaussien,
– un retard à la transmission infini.

En pratique, l'égalisation des canaux et le codage correcteur d'erreurs permettent de se rapprocher de cette limite avec un retard de transmission fini.

1.14 LES REPRÉSENTATIONS BINAIRES

Il existe diverses façons d'établir la correspondance entre l'ensemble des amplitudes quantifiées et l'ensemble des nombres binaires qui doivent les représenter. Les signaux à coder ayant des amplitudes en général positives et négatives, les représentations préférées sont celles qui conservent l'information de signe. Les plus courantes pour les codages à échelon constant sont les suivantes :

– signe et valeur absolue
– binaire décentré
– complément à 1
– complément à 2.

Les définitions et particularités de ces représentations sont données dans la référence [12], le tableau 1.3 les définit pour 3 bits.

Les représentations en signe et valeur absolue et en binaire décentré sont les plus commodes pour la conversion Analogique/Numérique ; les deux autres sont surtout utilisées dans les circuits de calcul numérique ; elles sont présentées en détail par la suite.

Comme indiqué au paragraphe 1.11, qui décrit une représentation particulière importante du domaine des télécommunications, le codage non linéaire permet d'augmenter considérablement la dynamique. Les machines de traitement, et en particulier celles qui sont à usage général, utilisent souvent des représentations dites à virgule flottante où chaque nombre comporte trois parties : le bit de signe, la mantisse et l'exposant. La mantisse représente une partie fractionnaire et l'exposant la puissance d'un nombre de base ; par exemple, en base 10 : $+ 0{,}719 \times 10^5$.

Tableau 1.3. – REPRÉSENTATIONS BINAIRES POUR CODAGE LINÉAIRE.

Nombre	Signe et valeur absolue	Binaire décentré	Complément à 1	Complément à 2
+ 3	0 1 1	1 1 1	0 1 1	0 1 1
+ 2	0 1 0	1 1 0	0 1 0	0 1 0
+ 1	0 0 1	1 0 1	0 0 1	0 0 1
+ 0	0 0 0	1 0 0	0 0 0	0 0 0
– 0	1 0 0	–	1 1 1	–
– 1	1 0 1	0 1 1	1 1 0	1 1 1
– 2	1 1 0	0 1 0	1 0 1	1 1 0
– 3	1 1 1	0 0 1	1 0 0	1 0 1
		(0 0 0)		(1 0 0)

L'extension de la dynamique provient de l'effet multiplicatif introduit par l'exposant. Ainsi en base 2, pour un nombre à 6 bits d'exposant et 16 bits de mantisse, la dynamique correspond à $2^{64} \times 2^{16} = 2^{80} \simeq 10^{24}$, soit 24 chiffres décimaux. Un grain supplémentaire est obtenu en prenant une base qui est elle même une puissance de deux, comme 8 ou 16, correspondant aux numérations octale ou hexadécimale respectivement.

La présentation à virgule flottante entraîne cependant une complication des opérations arithmétiques et des circuits.

Les nombres issus du codage se présentent, suivant la technique utilisée, soit sous forme parallèle, c'est-à-dire que les N bits sont disponibles sur N points de connexion en même temps, soit sous forme série, c'est-à-dire que les N bits apparaissent successivement sur le même point de connexion, le signe d'abord et ensuite les bits de poids décroissants. La référence [13] décrit les principales techniques de conversion Analogique/Numérique.

ANNEXE 1 : La fonction I(x)

Suite des valeurs : $\mathrm{I}(n) = \dfrac{\sin\left(\pi \frac{n}{20}\right)}{\pi \frac{n}{20}}$ pour $0 \leqslant n \leqslant 159$ avec $n = k + 20\,\mathrm{N}$.

k	N = 0	N = 1	N = 2	N = 3	N = 4	N = 5	N = 6	N = 7
0	1	0	0	0	0	0	0	0
1	0,99589	– 0,04742	0,02429	– 0,01633	0,01229	– 0,00986	0,00823	– 0,00706
2	0,98363	– 0,08942	0,04684	– 0,03173	0,02399	– 0,01929	0,01613	– 0,01385
3	0,96340	– 0,12566	0,06721	– 0,04588	0,03482	– 0,02806	0,02350	– 0,02021
4	0,93549	– 0,15591	0,08504	– 0,05847	0,04455	– 0,03598	0,03018	– 0,02599
5	0,90032	– 0,18006	0,10004	– 0,06926	0,05296	– 0,04287	0,03601	– 0,03105
6	0,85839	– 0,19809	0,11196	– 0,07804	0,05989	– 0,04850	0,04088	– 0,03528
7	0,81033	– 0,21009	0,12069	– 0,08466	0,06520	– 0,05301	0,04466	– 0,03859
8	0,75683	– 0,21624	0,12614	– 0,08904	0,06880	– 0,05606	0,04730	– 0,04091
9	0,69865	– 0,21682	0,12832	– 0,09113	0,07065	– 0,05769	0,04874	– 0,04220
10	0,63662	– 0,21221	0,12732	– 0,09095	0,07074	– 0,05787	0,04897	– 0,04244
11	0,57162	– 0,20283	0,12329	– 0,08856	0,06910	– 0,05665	0,04800	– 0,04164
12	0,50455	– 0,18921	0,11643	– 0,08409	0,06581	– 0,05406	0,04587	– 0,03983
13	0,43633	– 0,17189	0,10702	– 0,07770	0,06099	– 0,05020	0,04265	– 0,03707
14	0,36788	– 0,15148	0,09538	– 0,06960	0,05479	– 0,04518	0,03844	– 0,03344
15	0,30011	– 0,12862	0,08185	– 0,06002	0,04739	– 0,03914	0,03335	– 0,02904
16	0,23387	– 0,10394	0,06682	– 0,04924	0,03898	– 0,03298	0,02751	– 0,02399
17	0,17001	– 0,07811	0,05071	– 0,03753	0,02980	– 0,02470	0,02110	– 0,01841
18	0,10929	– 0,05177	0,03392	– 0,02522	0,02007	– 0,01667	0,01426	– 0,01245
19	0,05242	– 0,02554	0,01688	– 0,01261	0,01006	– 0,00837	0,00716	– 0,00626

ANNEXE 2 : La loi Normale Réduite

$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$		$P = \frac{2}{\sqrt{2\pi}} \int_{\lambda}^{\infty} e^{-\frac{x^2}{2}} dx$		$P = \frac{2}{\sqrt{2\pi}} \int_{\lambda}^{\infty} e^{-\frac{x^2}{2}} dx$	
x	$10^5 . f(x)$	100 P	λ	λ	100 P
0	39894	100	0	0	100
0,2	39104	95	0,0627	0,2	84,148
0,4	36827	90	0,1257	0,4	68,916
0,6	33322	85	0,1891	0,6	54,851
0,8	28969	80	0,2533	0,8	42,371
1	24197	75	0,3186	1	31,731
1,2	19419	70	0,3853	1,2	23,014
1,4	14973	65	0,4538	1,4	16,151
1,6	11092	60	0,5244	1,6	10,960
1,8	7895	55	0,5978	1,8	7,186
2	5399	50	0,6745	2	4,550
2,2	3547	45	0,7554	2,2	2,781
2,4	2239	40	0,8416	2,4	1,640
2,6	1358	35	0,9346	2,6	0,932
2,8	792	30	1,0364	2,8	0,511
3	443	25	1,1503	3	0,270
3,2	238	20	1,2816	3,2	0,137
3,4	123	15	1,4395	3,4	0,067
3,6	61	10	1,6449	3,6	0,032
3,8	29	5	1,9600	3,8	0,014
4	13	1	2,5758	4	0,006
4,2	5,9	0,1	3,2905	4,5	0,00068
4,4	2,5	0,01	3,8906	5	0,000057
4,6	1	0,001	4,4172	5,5	0,000004
		0,0001	4,8916		
		0,00001	5,3267		

Approximation pour les grandes valeurs du paramètre λ :

$$P \approx \frac{3}{4} \frac{1}{\lambda} e^{-\frac{\lambda}{2}}$$

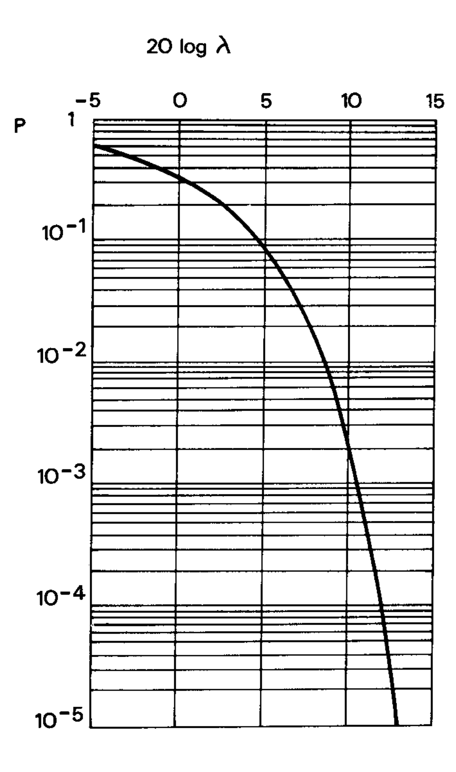


FIG. 1.19. *Loi Normale Réduite. Courbe donnant* P *en fonction de 20 log* λ

BIBLIOGRAPHIE

[1] L. SCHWARTZ – *Méthodes mathématiques pour les Sciences Physiques.* Éd. Hermann, 1961.

[2] E. ROUBINE – *Introduction à la théorie de la communication.* Éd. Masson, 2e édition, 1979.

[3] B. PICINBONO – Signaux aléatoires. Tomes 1 et 2, Éd. Dunod, 1994.

[4] J. R. RICE – *The Approximation of functions*. Reading Mass. Addison-Wesley, 1964.

[5] B. PICINBONO – *Éléments de Théorie du Signal*. Éd. Dunod, 1977.

[6] W. PETERSON – *Error Correcting Codes*. MIT Press, 1972.

[7] W. B. BENNET – *Spectra of quantized signals.* – The Bell System Technical Journal (BSTJ), Juillet 1948.

[8] CCITT – Réseaux numériques. Systèmes de transmission et équipements de multiplexage, Vol. III-3, Genève, Suisse, 1981.

[9] J. MAX – *Quantizing for Minimum Distortion.* IRE Transactions on Information Theory, Vol. IT-6, pp. 7-12, March 1960.

[10] M. D. PAEZ and T. H. GLISSON – *Minimum mean-Squared error quantization in Speech PCM and DPCM systems.* IEEE Trans. on Communications, Vol. COM-20, pp. 225-230, April 1972.

[11] C. SHANNON – *Communication in the presence of noise*, Proceedings of I.R.E., Vol. 37, pp. 10-21, Janv. 1949 (réimprimé dans : Proceedings of the IEEE, Sept. 1984 et Feb. 1998).

[12] P. DEBRAINE – *Machines de traitement de l'information*. Tome I, Éd. Masson, 1972.

[13] R. VAN DE PLASSCHE – « Integrated Analog to Digital and Digital-to-Analog Converters » Kluwer Academic Publishers, 1994.

EXERCICES

1 Soit le développement en série de Fourier de la Fonction $i(t)$ périodique et de période T, nulle sur toute la période, sauf dans l'intervalle $-\frac{\tau}{2} \leqslant t \leqslant \frac{\tau}{2}$ où elle prend la la valeur 1.

Donner la valeur des coefficients pour $\tau = \frac{T}{2}$ et $\tau = \frac{T}{3}$.

Vérifier que le développement conduit bien à : $i(0) = 1$ et tracer la fonction quand le développement est limité à 5 termes.

2 Analyser l'échantillonnage à la fréquence *fe* du signal $s(t) = \sin(\pi f_e t + \varphi)$ quand φ varie de 0 à $\frac{\pi}{2}$.

Examiner la reconstitution de ce signal à partir des échantillons.

3 Calculer la distorsion d'amplitude apportée à un signal restitué par des impulsions dont la largeur est égale à la moitié de la période d'échantillonnage.

4 Échantillonnage passe-bande d'un signal occupant la bande de fréquence $[f_1 - f_2]$. Quelles sont les conditions à imposer à la fréquence f_1 pour que ce signal puisse être échantillonné directement à une fréquence comprise entre f_2 et $2f_2$.

5 Analyser l'échantillonnage du signal suivant :

$$s_i(t) = \sum_{n=1}^{3} \sin\left(2\pi \frac{n}{8} \frac{t}{T}\right)$$

et le comparer à celui du signal :

$$s_r(t) = \sum_{n=1}^{3} \cos\left(2\pi \frac{n}{8} \frac{t}{T}\right)$$

Montrer par une étude des spectres que l'ensemble des deux suites d'échantillons constitue l'échantillonnage d'un signal complexe.

6 Soit $s(t)$ le signal défini par l'égalité :

$$s(t) = 1 + 2 \sum_{k=1}^{3} \cos\left(2\pi \frac{kt}{8} + \varphi_k\right) + \cos(\pi t + \varphi_4)$$

Ce signal est échantillonné avec la période T = 1. Quelle est la valeur maximale que peut prendre $s(n)$ avec n entier ? Montrer qu'il existe un ensemble de valeurs φ_k ($k = 1, 2, 3, 4$) qui minimisent la valeur maximale des $s(n)$. Peut-on généraliser cette propriété ?

7 Un synthétiseur de fréquence numérique est construit à partir d'une mémoire morte de 16*k*bits ayant un temps d'accès de 500 *ns.* Sachant que les nombres qui représentent les échantillons de signaux sinusoïdaux comptent 8 bits, quelles sont les caractéristiques du synthétiseur, gamme et pas de fréquence, qui peuvent être obtenues ?

8 Quelle est la loi de probabilité des amplitudes du signal sinusoïdal suivant :

$$s(t) = A \cos\left(2\pi \frac{t}{T}\right)$$

Calculer la fonction d'autocorrélation correspondante.

Donner la fonction d'autocorrélation d'un signal aléatoire gaussien stationnaire dont le spectre a une répartition uniforme dans la bande de fréquence $[f_1, f_2]$.

9 Calculer le spectre d'une suite d'impulsions de largeur T/2, séparées de T, chaque impulsion ayant la probabilité p d'apparaître. Examiner en particulier le cas où $p = 1/2$.

Que devient ce spectre si ces impulsions constituent une séquence pseudo-aléatoire de longueur $2^4 - 1 = 15$ engendrée par un registre à 4 bits, suivant le polynôme $g(x) = x^4 + x + 1$?

10 Un signal sinusoïdal à la fréquence 1 050 Hz est échantillonné à 8 kHz et codé à 10 bits. Quelle est la valeur du rapport signal à bruit maximal ? Quel est le niveau par rapport au signal du bruit de quantification mesuré dans la bande de fréquence 300-500 Hz ? Même question si la fréquence d'échantillonnage est portée à 16 kHz.

11 Le signal $\sin\left(2\pi \frac{t}{8} + \varphi\right)$ avec $0 \leqslant \varphi \leqslant \frac{\pi}{2}$ est échantillonné avec la période T = 1 et codé à 5 bits.

Dans le cas où $\varphi = 0$ calculer la puissance et le spectre du bruit de quantification.

Comment évolue ce spectre en fonction de la phase φ?

12 Soit une échelle de codage où l'échelon a la valeur q. Étudier la quantification du signal $s_1(t) = \alpha \,.\, q \,.\sin(\omega_1, t)$ pour $-1 \leqslant \alpha \leqslant 1$, en fonction du centrage de la caractéristique de quantification. Donner l'enveloppe du signal restitué après décodage et filtrage étroit autour de la fréquence ω_1.

Au signal $s_1(t)$ on superpose le signal $s_2(t) = 10.q.\sin \omega_2 t$. Étudier l'enveloppe du signal restitué dans ces conditions.

13 Soit à coder un signal gaussien. Combien de bits faut-il pour que le rapport signal à bruit de quantification soit meilleur que 50 dB ? Peut-on réduire ce nombre si l'on admet un écrêtage pendant 1 % du temps.

14 Le signal $s(t) = A \sin(2\pi \ .810.t)$ est codé à 8 bits. L'échelon de quantification ayant la valeur q, tracer la courbe donnant le rapport signal à bruit de quantification en fonction de l'amplitude A lorsque cette amplitude varie de q à $2^7.q$. Même question lorque le codage est du type non linéaire suivant la loi à 13 segments.

15 Calculer les limites des plages d'amplitude élémentaires pour le codage optimal à 2 bits d'un signal gaussien de puissance unité.

Chapitre 2

La transformation de Fourier discrète

La transformation de Fourier Discrète s'introduit quand il s'agit de calculer la transformée de Fourier d'une fonction à l'aide d'un calculateur numérique. En effet un tel opérateur ne peut traiter que des nombres et de plus en quantité limitée par la taille de sa mémoire. Il s'en suit que la transformée de Fourier :

$$S(f) = \int_{-\infty}^{\infty} s(t)\, e^{-j2\pi f t}\, dt$$

doit être adaptée, d'une part en remplaçant le signal $s(t)$ par des nombres $s(nT)$ qui représentent un échantillonnage de ce signal et d'autre part en limitant l'ensemble des nombres sur lesquels portent les calculs à une valeur finie N. Le calcul fournit alors des nombres $S^*(f)$ définis par :

$$S^*(f) = \sum_{n=0}^{N-1} s(nT)\, e^{-j2\pi f nT}$$

Comme le calculateur est limité dans sa puissance de calcul, il ne peut fournir ces résultats que pour un nombre limité de valeurs de la fréquence f, qu'il est naturel de choisir multiples d'un certain pas de fréquence Δf. Alors :

$$S^*(k\Delta f) = \sum_{n=0}^{N-1} s(nT)\, e^{-j2\pi nk\Delta f T}$$

Les conditions dans lesquelles les valeurs calculées constituent une bonne approximation des valeurs recherchées sont étudiées par la suite. Un choix simplificateur intéressant consiste à prendre : $\Delta f = \dfrac{1}{NT}$. On peut alors vérifier qu'il existe seulement N valeurs différentes dans la suite des $S^*(k/NT)$, qui est une suite périodique et de période N puisque

$$S^*[(k+N)/NT] = S^*(k/NT)$$

D'autre part la transformée ainsi calculée se présente sous la forme de valeurs discrètes et d'après le paragraphe I.6 sur l'échantillonnage en fréquence, cette propriété est caractéristique du spectre des fonctions périodiques. On peut donc considérer que la suite des $S^*(k/NT)$ est obtenue par transformation de Fourier de la suite des $s(nT)$ qui est une suite périodique et de période NT.

La transformation de Fourier discrète (T.F.D.) et la transformée inverse établissent des relations entre ces deux suites périodiques.

La définition, les propriétés, les techniques de calcul et les applications de la T.F.D. ont fait l'objet de nombreux articles et ouvrages, parmi lesquels on peut citer les références [1, 2, 3, 4].

2.1 DÉFINITION ET PROPRIÉTÉS DE LA TFD

Soit deux suites de nombres complexes $x(n)$ et $X(k)$, périodiques et de période N. La transformée de Fourier Discrète et la transformée inverse établissent entre ces deux suites les relations suivantes respectivement :

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-j2\pi \frac{nk}{N}} \tag{2.1}$$

$$x(n) = \sum_{k=0}^{N-1} X(k) e^{j2\pi \frac{kn}{N}} \tag{2.2}$$

La position du facteur d'échelle 1/N est choisie pour que les $X(k)$ soient les coefficients du développement en série de Fourier de la suite $x(n)$. Cette transformation possède les propriétés suivantes :

– **Linéarité :** si $x(n)$ et $y(n)$ sont deux suites de même période, dont les transformées sont $X(k)$ et $Y(k)$ respectivement, la suite de même période $v(n) = x(n) + \lambda y(n)$ où λ est un scalaire a pour transformée :

$$V(k) = X(k) + \lambda . Y(k)$$

– **Une translation des $x(n)$ entraîne une rotation de phase des $X(k)$:** En effet calculons la transformée $X_{n_0}(k)$ de la suite $x(n - n_0)$.

$$X_{n_0}(k) = \sum_{n=0}^{N-1} x(n-n_0) e^{-j2\pi \frac{nk}{N}} = X(k) e^{-j2\pi n_0 k/N}$$

Une translation de $x(n)$ de n_0 entraîne sur $X(k)$ une rotation de la phase d'un angle égal à : $2\pi \frac{n_0 k}{N}$.

– **Symétrie :** si la suite $x(n)$ est réelle les nombres $X(k)$ et $X(N - k)$ sont complexes conjugués :

$$\overline{X}(N-k) = \sum_{n=0}^{N} x(n) e^{j2\pi \frac{n(N-k)}{N}} = X(k)$$

Si la suite $x(n)$ est réelle et paire il en est de même de la suite $X(k)$. En effet, si $x(N-n) = x(n)$, il vient par exemple pour $N = 2P + 1$:

$$X(N-k) = x(0) + 2\sum_{n=1}^{P} x(n)\cos\left(2\pi\frac{nk}{N}\right) = X(k)$$

Si la suite $x(n)$ est réelle et impaire la suite $X(k)$ est purement imaginaire. Dans ce cas : $x(N-n) = -x(n)$ et $x(0) = x(N) = 0$. Par exemple pour $N = 2P + 1$, il vient :

$$X(k) = -2j\sum_{n=1}^{P} x(n)\sin\left(2\pi\frac{nk}{N}\right) = -X(N-k)$$

On peut remarquer que $X(0) = X(N) = 0$.

Tout signal réel pouvant toujours se décomposer en une partie paire et une partie impaire, ces deux dernières propriétés de symétrie sont importantes.

– **Convolution circulaire :** la transformée d'un produit de convolution est égale au produit des transformées.

Si $x(n)$ et $h(n)$ sont deux suites de période N, la convolution circulaire $y(n)$ peut être définie par l'équation :

$$y(n) = \sum_{l=0}^{N-1} x(l)h(n-l) \tag{2.3}$$

C'est une suite qui possède la même période N. Sa transformée s'écrit :

$$Y(k) = \sum_{n=0}^{N-1}\left[\sum_{l=0}^{N-1} x(l)h(n-l)\right]e^{-j2\pi\frac{nk}{N}} = \sum_{l=0}^{N-1} x(l)\left[\sum_{n=0}^{N-1} h(n-l)e^{-j2\pi(n-l)k/N}\right].e^{-j2\pi\frac{lk}{N}}$$

$$Y(k) = \left(\sum_{n=0}^{N-1} h(n-l)e^{-j2\pi\frac{(n-l)k}{N}}\right)\left(\sum_{l=0}^{N-1} x(l)e^{-j2\pi\frac{lk}{N}}\right) = H(k).X(k) \tag{2.4}$$

C'est une propriété majeure de la transformation de Fourier Discrète. Une application directe en sera donnée ultérieurement.

– **Égalité de Parseval :** elle exprime que la puissance du signal est égale à la somme des puissances de ses harmoniques. En effet :

$$\frac{1}{N}\sum_{n=0}^{N-1} x(n)\bar{x}(n) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\sum_{k=0}^{N-1}\overline{X}(k)e^{-j2\pi\frac{kn}{N}}$$

$$\frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2 = \sum_{k=0}^{N-1}\overline{X}(k)\,\frac{1}{N}\sum_{n=0}^{N-1} x(n)e^{-j2\pi\frac{kn}{N}}$$

$$\frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2 = \sum_{k=0}^{N-1}|X(k)|^2 \tag{2.5}$$

– **Relation avec la série de Fourier :** avec la position du facteur d'échelle 1/N choisie dans la définition (2.1) de la TFD, les valeurs $X(k)$ représentent, aux repliements de spectre près, les coefficients du développement en série de Fourier du signal périodique, quand ce signal ne présente pas de discontinuité. Si ce n'est pas le cas, des différences importantes apparaissent. En effet, on montre qu'à un point de discontinuité t_0 de la fonction périodique $x(t)$, le développement en série de Fourier de $x(t)$ prend une valeur égale à la moyenne des limites à gauche et à droite de $x(t)$ quand t tend vers t_0. Par contre, la TFD inverse restitue exactement les échantillons du signal d'origine et par suite les valeurs $X(k)$ comprennent la TFD de la distribution des discontinuités avec le signe inverse et le facteur 1/2.

Exemple : soit le signal triangulaire :

$$\begin{aligned} x(t) &= t;\ 0 \leq t < 1 \\ x(t+1) &= x(t) \end{aligned}$$

Coefficients du développement en série de Fourier :

$$\begin{aligned} &C_0 = 1/2 \\ &C_n = j/2\pi n;\ n\ \textit{entier};\ n \neq 0 \end{aligned}$$

La discontinuité à l'origine a une amplitude égale à 1 et la transformée de Fourier discrète d'ordre N donne les valeurs suivantes :

$$X(k) = -\frac{1}{2N} + X'(k) \text{ avec } X'(k) \approx C_k$$

Cette particularité de la TFD par rapport à la série de Fourier est un effet indésirable quand on cherche un développement avec des coefficients ayant les valeurs les plus faibles possibles, comme dans la compression des signaux.

Cependant la propriété la plus importante de la Transformation de Fourier Discrète réside probablement dans le fait qu'elle se prête à des techniques de calcul efficaces, qui lui ont donné une place prépondérante en traitement numérique du signal.

2.2 LA TRANSFORMATION DE FOURIER RAPIDE

Les équations de définition de la TFD fournissent une relation entre deux ensembles de N nombres complexes, qui s'écrit d'une manière commode sous une forme matricielle, en posant :

$$W = e^{-j\frac{2\pi}{N}} \tag{2.6}$$

Les affixes des nombres W^n, appelés coefficients de la T.F.D. se trouvent sur le cercle unité comme le montre la figure 2.1. Ce sont les racines de l'équation $Z^N - 1 = 0$ ou racines Nièmes de l'unité.

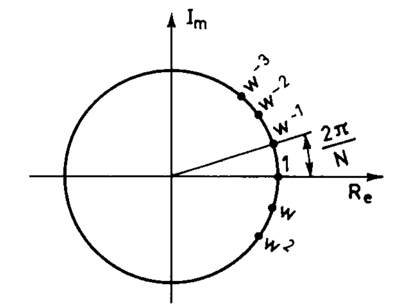


FIG. 2.1. *Affixes des coefficients de la TFD*

L'équation matricielle est la suivante pour la transformée directe.

$$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ \vdots \\ X_{N-1} \end{bmatrix} = \frac{1}{N} \begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & W & W^2 & W^3 & \dots & W^{N-1} \\ 1 & W^2 & W^4 & W^6 & \dots & W^{2(N-1)} \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & W^{(N-1)} & W^{2(N-1)} & & \dots & W^{(N-1)(N'-1)} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{N-1} \end{bmatrix}$$

Pour la transformée inverse, il suffit de retirer le scalaire 1/N et de changer W^n en W^{-n}.

La matrice carrée d'ordre N désignée par T_N présente des particularités évidentes, les lignes et les colonnes de même indice ont les mêmes éléments et ces éléments sont des puissances d'un nombre de base W tel que $W^N = 1$. Des simplifications importantes peuvent être envisagées dans ces conditions, conduisant à des algorithmes de calcul rapide. Quand la TFD est calculée à l'aide de tels algorithmes on dit que l'on effectue une Transformation de Fourier Rapide (TFR).

Un cas très intéressant est celui où N est une puissance de deux car il conduit à des algorithmes peu complexes qui sont particulièrement efficaces. Ces algorithmes sont basés sur une décomposition de la suite à transformer en sous-suites entrelacées. Le cas de l'entrelacement dans le temps va être considéré d'abord.

2.2.1 TFR avec entrelacement temporel

La suite d'éléments $x(n)$ peut être décomposée en deux suites entrelacées, celle des éléments d'indice pair et celle des éléments d'indice impair. Calculons, en faisant apparaître cette décomposition, les N/2 premiers éléments de l'ensemble des $X(k)$:

$$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ \vdots \\ X_{N/2-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & W^2 & \dots & W^{2(N/2-1)} \\ 1 & W^4 & \dots & W^{4(N/2-1)} \\ \vdots & \vdots & & \vdots \\ 1 & W^{2(N/2-1)} & \dots & W^{2(N/2-1)(N/2-1)} \end{bmatrix} \begin{bmatrix} x_0 \\ x_2 \\ x_4 \\ \vdots \\ x_{2(N/2-1)} \end{bmatrix}$$

$$+\begin{bmatrix} 1 & 1 & \dots & 1 \\ W & W^3 & \dots & W^{N-1)} \\ W^2 & W^6 & \dots & W^{2(N-1)} \\ \vdots & \vdots & & \vdots \\ W^{N/2-1)} & W^{3(N/2-1)} & \dots & W^{(N/2-1)(N-1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_3 \\ x_5 \\ \vdots \\ x_{N-1} \end{bmatrix}$$

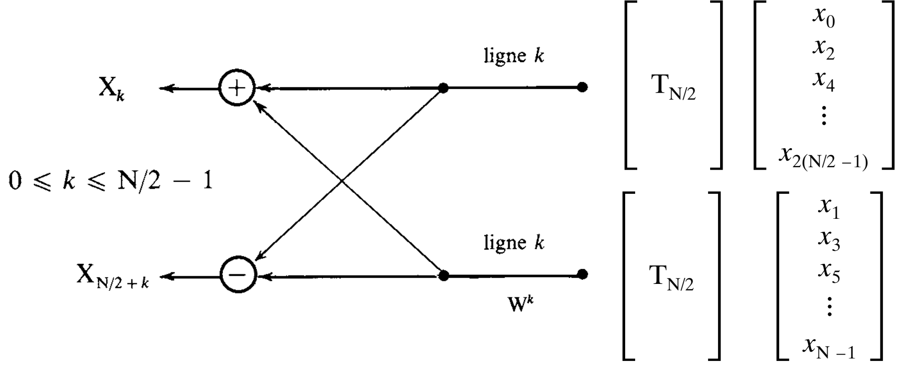
En désignant par $T_{N/2}$ la matrice qui vient en facteur du vecteur colonne des éléments d'indice pair et en décomposant la matrice facteur du vecteur des éléments d'indice impair en un produit d'une matrice diagonale par la matrice $T_{N/2}$, on obtient :

$$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ \vdots \\ X_{N/2-1} \end{bmatrix} = T_{N/2} \begin{bmatrix} x_0 \\ x_2 \\ x_4 \\ \vdots \\ x_{2(N/2-1)} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & W & 0 & \dots & 0 \\ 0 & 0 & W^2 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \dots & & W^{N/2-1} \end{bmatrix} T_{N/2} \begin{bmatrix} x_1 \\ x_3 \\ x_5 \\ \vdots \\ x_{N-1} \end{bmatrix}$$

Et pour les N/2 derniers éléments de l'ensemble des $X(k)$, compte tenu du fait que $W^N = 1$:

$$\begin{bmatrix} X_{N/2} \\ X_{N/2+1} \\ X_{N/2+2} \\ \vdots \\ X_{N-1} \end{bmatrix} = T_{N/2} \begin{bmatrix} x_0 \\ x_2 \\ x_4 \\ \vdots \\ x_{2(N/2-1)} \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & W & 0 & \dots & 0 \\ 0 & 0 & W^2 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & W^{N/2-1} \end{bmatrix} T_{N/2} \begin{bmatrix} x_1 \\ x_3 \\ x_5 \\ \vdots \\ x_{N-1} \end{bmatrix}$$

Il apparaît que le calcul de $X(k)$ et $X(k + N/2)$ pour $0 \leqslant k \leqslant N/2 - 1$ met en œuvre les mêmes calculs avec seulement un changement de signe dans la somme finale. D'où le diagramme suivant :



Ce diagramme montre que le calcul d'une Transformée de Fourier d'ordre N revient au calcul de deux transformées d'ordre N/2 auquel s'ajoutent N/2 multiplications complexes. Par itération en un nombre d'étapes égal à log 2 (N) – 1 = log 2 (N/2), on aboutit aux transformées d'ordre 2, dont la matrice s'écrit :

$$T_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

et qui ne demandent pas de multiplications.

Comme chaque étape comporte N/2 multiplications complexes, l'ensemble de la transformation demande un nombre de multiplications complexes M_c qui s'écrit :

$$M_c = N/2 \log 2\,(N/2) \tag{2.7}$$

et un nombre d'additions complexes A*c* tel que :

$$A_c = N \log 2\,(N) \tag{2.8}$$

En réalité le nombre de multiplications complexes peut encore être réduit parce que certaines puissances de W présentent des particularités; $W^0 = 1$ et $W^{N/4} = -j$ ne demandent pas de multiplications complexes;

$$W^{N/8} = \frac{\sqrt{2}}{2}\,(1 - j) \qquad \text{et} \qquad W^{3N/8} = \frac{\sqrt{2}}{2}\,(-1 - j)$$

ne demandent qu'une demi-multiplication complexe chacune. Dans la première étape on peut ainsi économiser 3 multiplications, dans l'avant dernière étape 3. N/8 et dans la dernière 2. N/4. Le gain dans l'ensemble des étapes s'élève à 5N/4 – 3 et le nombre minimum de multiplications complexes est donné par :

$$m_c = N/2\left[\log 2\,(N/2) - \frac{5}{2}\right] + 3 \tag{2.9}$$

Il faut toutefois noter que toutes ces réductions de calcul ne sont pas toujours faciles à exploiter, aussi bien en logiciel qu'en matériel.

La transformée d'ordre 4 a pour matrice :

$$T_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -j & -1 & +j \\ 1 & -1 & 1 & -1 \\ 1 & +j & -1 & -j \end{bmatrix} \tag{2.10}$$

Son diagramme est donné par la figure 2.2.

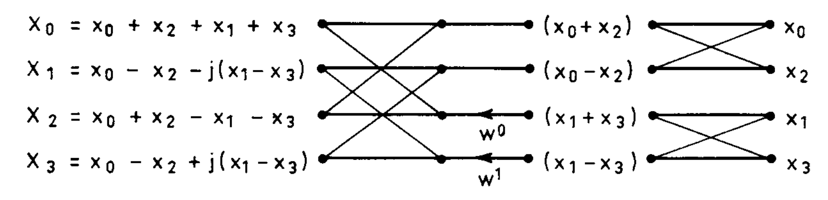


FIG. 2.2. *Transformée d'ordre 4 avec entrelacement temporel*

Par convention les flèches représentent les multiplications, les points à gauche des croisillons élémentaires représentent, le point supérieur une addition, le point inférieur une soustraction. La transformée d'ordre 8 est représentée sur la figure 2.3.

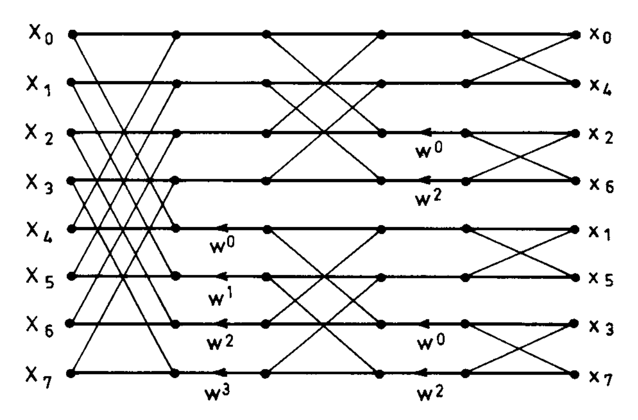


FIG. 2.3. *Transformée d'ordre 8 avec entrelacement temporel*

On peut remarquer que dans ce traitement les $X(k)$ apparaissent dans l'ordre naturel des indices alors que les $x(n)$ se présentent dans un ordre permuté. Cette permutation est due aux entrelacements successifs et se traduit par un retournement ou inversion de la représentation binaire des indices. Par exemple pour $N = 8$:

à :	x_0 (0 0 0)	correspond :	x_0 (0 0 0)
	x_1 (0 0 1)	–	x_4 (1 0 0)
	x_2 (0 1 0)	–	x_2 (0 1 0)
	x_3 (0 1 1)	–	x_6 (1 1 0)
	x_4 (1 0 0)	–	x_1 (0 0 1)
	x_5 (1 0 1)	–	x_5 (1 0 1)
	x_6 (1 1 0)	–	x_3 (0 1 1)
	x_7 (1 1 1)	–	x_7 (1 1 1)

La quantité de mémoires de données nécessaire pour calculer une transformée d'ordre N est de N positions complexes. En effet les calculs sont faits sur des couples de variables qui subissent l'opération représentée par un croisillon et conservent leur position dans l'ensemble des variables à la fin de cette opération, comme l'indiquent clairement les diagrammes. D'autre part, la transformée inverse est obtenue en changeant simplement le signe des exposants de W. On introduit le facteur 1/N, par exemple en multipliant par 1/2 les résultats des additions et soustractions effectuées dans les croisillons ce qui permet de conserver le cadrage des nombres dans les mémoires.

On peut aussi faire les calculs sans inversion binaire, il faut alors 2N mémoires de données, car les couples de nombres changent de position à chaque étage.

Le type d'entrelacement qui vient d'être étudié peut aussi être opéré sur les $X(k)$; un algorithme similaire est alors obtenu.

2.2.2 TFR avec entrelacement fréquentiel

La suite des éléments $X(k)$ peut être décomposée en 2 suites entrelacées, celle des éléments d'indice pair et celle des éléments d'indice impair. Pour les éléments

d'indice pair, en tenant compte du fait que $W^N = 1$, il vient après une mise en facteur élémentaire :

$$\begin{bmatrix} X_0 \\ X_2 \\ X_4 \\ \vdots \\ X_{2(N/2-1)} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & W^2 & W^4 & \dots & W^{2(N/2-1)} \\ 1 & W^4 & W^8 & \dots & W^{4(N/2-1)} \\ \vdots & \vdots & & & \vdots \\ 1 & W^{2(N/2-1)} & & \cdots & W^{2(N/2-1)(N/2-1)} \end{bmatrix} \begin{bmatrix} x_0 + x_{N/2} \\ x_1 + x_{N/2+1} \\ x_2 + x_{N/2+2} \\ \vdots \\ x_{N/2-1} + x_{N-1} \end{bmatrix}$$

Et pour les éléments d'indice impair, après mise en facteur :

$$\begin{bmatrix} X_1 \\ X_3 \\ X_5 \\ \vdots \\ X_{N-1} \end{bmatrix} = \begin{bmatrix} 1 & W & W^2 & \dots & W^{N/2-1} \\ 1 & W^3 & W^6 & \dots & W^{3(N/2-1)} \\ 1 & W^5 & W^{10} & \dots & W^{5(N/2-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & W^{N-1} & W^{2(N-1)} & \cdots & W^{(N-1)(N/2-1)} \end{bmatrix} \begin{bmatrix} x_0 - x_{N/2} \\ x_1 - x_{N/2+1} \\ x_2 - x_{N/2+2} \\ \vdots \\ x_{N/2-1} - x_{N-1} \end{bmatrix}$$

Dans ce cas la matrice carrée obtenue est égale au produit de la matrice carrée $T_{N/2}$ obtenue pour les éléments d'indice pair par la matrice diagonale dont les éléments sont les puissances W^k avec $0 \leqslant k \leqslant N/2 - 1$. D'où :

$$\begin{bmatrix} X_1 \\ X_3 \\ X_5 \\ \vdots \\ X_{N-1} \end{bmatrix} = T_{N/2} \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & W & 0 & \dots & 0 \\ 0 & 0 & W^2 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & W^{N/2-1} \end{bmatrix} \begin{bmatrix} x_0 - x_{N/2} \\ x_1 - x_{N/2+1} \\ x_2 - x_{N/2+2} \\ \vdots \\ x_{N/2-1} - x_{N-1} \end{bmatrix}$$

Le calcul des éléments $X(k)$ d'indice pair et d'indice impair se fait avec la matrice carrée $T_{N/2}$ de la transformée d'ordre N/2 et le diagramme suivant est obtenu :

$$\begin{bmatrix} X_0 \\ X_2 \\ X_4 \\ \vdots \\ x_{2(N/2-1)} \end{bmatrix} = \begin{bmatrix} \\ T_{N/2} \\ \\ \end{bmatrix}$$

$$\begin{bmatrix} X_1 \\ X_3 \\ X_5 \\ \vdots \\ x_{N-1} \end{bmatrix} = \begin{bmatrix} \\ T_{N/2} \\ \\ \end{bmatrix}$$

ligne k — x_k

$0 \leqslant k \leqslant N/2 - 1$

ligne k — W^k — $x_{N/2+k}$

En adoptant la même notation pour les croisillons qu'au paragraphe précédent on établit des diagrammes semblables. La figure 2.4 montre le diagramme obtenu pour N = 8.

Dans l'entrelacement fréquentiel, le nombre de calculs est le même que dans l'entrelacement temporel; les nombres à transformer $x(n)$ apparaissent dans l'ordre naturel alors que les nombres transformés $X(k)$ sont permutés.

Les algorithmes qui ont été obtenus jusqu'à maintenant sont basés sur une décomposition de la transformée d'ordre N en transformées élémentaires d'ordre 2 qui ne nécessitent pas de multiplications. Ces algorithmes sont dits en base 2. Cependant d'autres transformées élémentaires peuvent être utilisées ; la plus intéressante est la transformée en base 4 qui s'appuie sur la matrice élémentaire T_4 et conduit aux algorithmes en base 4.

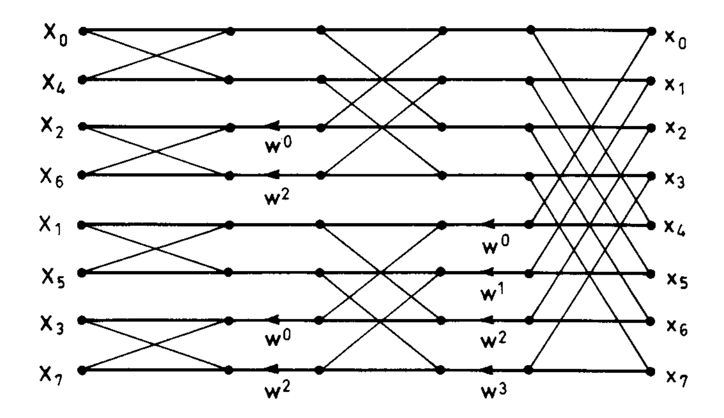


FIG. 2.4. *Transformée d'ordre 8 avec entrelacement fréquentiel*

2.2.3 Algorithme de TFR en base 4

Cet algorithme peut être utilisé lorsque N est une puissance de 4. La suite des nombres x(n) est décomposée en 4 suites entrelacées. Calculons les $\frac{N}{4}$ premiers $X(k)$ en mettant en évidence cette décomposition ; l'expression matricielle est alors la suivante, si $T_{N/4}$ désigne la matrice carrée de la transformée d'ordre N/4 et $D_i (i = 1, 2, 3)$ la matrice diagonale dont les éléments sont les puissances $W^{i.k}$ avec $0 \leqslant k \leqslant N/4 - 1$:

$$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ \vdots \\ X_{N/4-1} \end{bmatrix} = T_{N/4} \begin{bmatrix} x_0 \\ x_4 \\ x_8 \\ \vdots \\ x_{4(N/4-1)} \end{bmatrix} + D_1 T_{N/4} \begin{bmatrix} x_1 \\ x_5 \\ x_9 \\ \vdots \\ x_{N-3} \end{bmatrix} + D_2 T_{N/4} \begin{bmatrix} x_2 \\ x_6 \\ x_{10} \\ \vdots \\ x_{N-2} \end{bmatrix} + D_3 T_{N/4} \begin{bmatrix} x_3 \\ x_7 \\ x_{11} \\ \vdots \\ x_{N-1} \end{bmatrix}$$

Les N/4 termes $X(k)$ suivants sont donnés par :

$$\begin{bmatrix} X_{N/4} \\ X_{N/4+1} \\ \vdots \\ X_{N/2-1} \end{bmatrix} = T_{N/4} \begin{bmatrix} x_0 \\ x_4 \\ \vdots \\ x_{4(N/4-1)} \end{bmatrix} - jD_1T_{N/4} \begin{bmatrix} x_1 \\ x_5 \\ \vdots \\ x_{N-3} \end{bmatrix} - D_2T_{N/4} \begin{bmatrix} x_2 \\ x_6 \\ \vdots \\ x_{N-2} \end{bmatrix} + jD_3T_{N/4} \begin{bmatrix} x_3 \\ x_7 \\ \vdots \\ x_{N-1} \end{bmatrix}$$

Cette équation fait intervenir les mêmes calculs matriciels que la précédente avec en plus des multiplications par les éléments de la seconde ligne de la matrice T_4. On peut alors montrer que le calcul de la transformée aboutit au diagramme de la figure 2.5.

FIG. 2.5. *Transformée en base 4*

Une telle transformée s'effectue en $\log_4(N) - 1 = \log_4\left(\frac{N}{4}\right)$ étapes.

Chaque étape demande $3\,\frac{N}{4}$ multiplications complexes ce qui conduit au total au nombre de multiplications M_{c4} donné par :

$$M_{c4} = \frac{3}{4}\,N \log_4\left(\frac{N}{4}\right) \tag{2.11}$$

Le nombre d'additions complexes A_{c4} s'élève à :

$$A_{c4} = 2\,N \log_4(N) \tag{2.12}$$

Il apparaît que le nombre d'additions est le même en base 2 et en base 4, par contre pour les multiplications complexes, le calcul en base 4 apporte un gain supérieur à 25 %.

La figure 2.6. donne le diagramme complet pour N = 16.

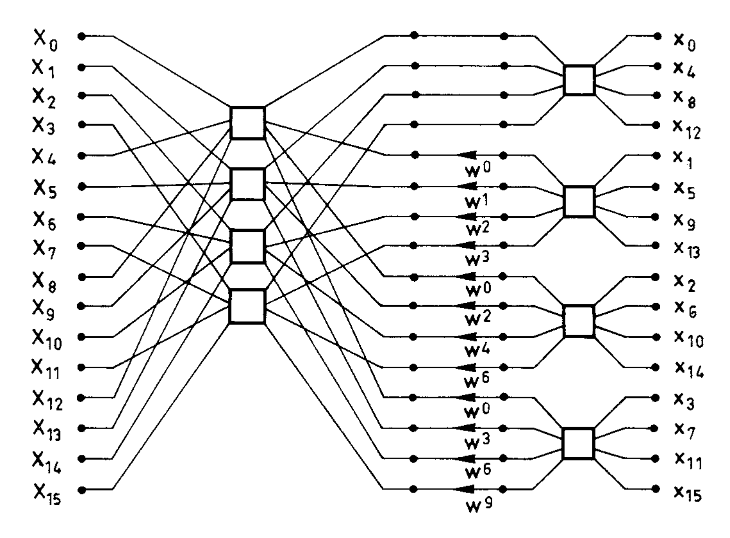


FIG. 2.6. *Transformée d'ordre 16 en base 4*

D'autres bases peuvent encore être envisagées, par exemple la base 8 ; dans ce cas, il y a des multiplications dans la matrice élémentaire et les gains par rapport à la base 4 sont minimes.

Des bases différentes peuvent également être combinées [5].

2.2.4 Algorithme de TFR en base double

Dans une transformée d'ordre N, la suite des valeurs transformées d'indices impairs peut être décomposée en deux suites, exprimées par les relations :

$$X(4k+1) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\,W^{n}\,W^{4kn}$$

et :

$$X(4k+3) = \frac{1}{N} \sum_{n=0}^{N-1} x(n)\, W^{3n} W^{4kn}$$

Compte tenu de la définition (2-6) de W les sommations peuvent aussi s'écrire :

$$X(4k+1) = \frac{1}{N} \sum_{n=0}^{N/4\ -1} \left[\left[x(n) - x\left(n + \frac{N}{2}\right)\right]\right.$$
$$\left. - j\left[x\left(n+\frac{N}{4}\right) - x\left(n + \frac{3N}{4}\right)\right]\right] W^{n} W^{4kn} \quad (2.13)$$

et :

$$X(4k+3) = \frac{1}{N} \sum_{n=0}^{N/4\ -1} \left[\left[x(n) - x\left(n + \frac{N}{2}\right)\right]\right.$$
$$\left. + j\left[x\left(n+\frac{N}{4}\right) - x\left(n + \frac{3N}{4}\right)\right]\right] W^{3n} W^{4kn} \quad (2.14)$$

La suite des valeurs transformées d'indices pairs s'écrit :

$$X(2k) = \frac{1}{N} \sum_{n=0}^{N/2\ -1} \left[x(n) + x\left(n + \frac{N}{2}\right)\right] W^{2nk} \quad (2.15)$$

Ces équations montrent que la première étape d'une transformée d'ordre N avec entrelacement temporel peut être remplacée par le calcul d'une transformée d'ordre N/2 et de deux transformées d'ordre N/4.

L'algorithme en base double est obtenu par applications successives de cette décomposition.

Pour l'entrelacement fréquentiel l'algorithme en base double est obtenu par la décomposition suivante :

$$X(k) = \sum_{n=0}^{N/2\ -1} x(2n)\, W^{2nk} + W^{k} \sum_{n=0}^{N/4\ -1} x(4n+1)\, W^{4nk}$$
$$+ W^{3k} \sum_{n=0}^{N/4\ -1} x(4n+3)\, W^{4nk} \quad (2.16)$$

Pour une transformée d'ordre N, le nombre de multiplications complexes $M_{c2/4}(N)$ est fourni par une récurrence déduite des relations ci-dessus :

$$M_{c2/4}(N) = M_{c2/4}\left(\frac{N}{2}\right) + 2\,M_{c2/4}\left(\frac{N}{4}\right) + \frac{N}{2} \quad (2.17)$$

avec comme valeurs initiales $M(2) = M(4) = 0$.

La valeur ainsi obtenue est légèrement inférieure à celle donnée par la base 4.

A titre de référence, il est intéressant de noter que le nombre minimal de multiplications complexes non triviales pour une transformée d'ordre N, avec $N = 2^m$, est égal à : $2^{m+1} - 2m^2 + 4m - 8$ [5].

Les algorithmes qui ont été présentés, entrelacement temporel et fréquentiel, en base deux et quatre, sont des éléments d'un ensemble d'algorithmes. Une présentation unifiée des algorithmes de TFR est donnée dans le chapitre suivant, elle permet de déterminer l'algorithme le plus approprié dans chaque application.

Dans les calculs, les opérations sont faites avec une précision limitée, ce qui amène certaines dégradations du signal.

2.3 DÉGRADATIONS DUES AUX LIMITATIONS DANS LE CALCUL

Les machines réelles apportent des limitations dans le calcul, qui sont dues aux opérateurs arithmétiques et aux mémoires. D'abord les coefficients sont souvent stockés dans une mémoire morte, avec un nombre de chiffres limité ; en fait le contenu de la mémoire représente une approximation des coefficients, en général obtenue par arrondi. Ensuite tout au long du calcul des arrondis sont opérés pour limiter le nombre de chiffres des nombres traités à la capacité des positions de mémoire ou des opérateurs arithmétiques. Ces deux types de limitations entraînent des dégradations qu'il est important d'analyser pour pouvoir déterminer avec précision le matériel strictement nécessaire à la mise en œuvre d'une Transformée ayant des performances spécifiées.

2.3.1 Effets de l'arrondi des coefficients

Les coefficients réellement utilisés par la machine représentent une approximation des coefficients théoriques, dont la valeur des parties réelle et imaginaire est comprise dans l'intervalle [– 1, + 1].

Une quantification à b_c bits, signe compris, entraîne sur le coefficient $e^{-j2\pi\frac{n}{N}}$ une erreur d'arrondi $\delta(n) = \delta_R(n) + j\delta_I(n)$ telle que l'on ait :

$$|\delta_R(n)| \leqslant 2^{-b_c} \quad \text{et} \quad |\delta_I(n)| \leqslant 2^{-b_c}$$

Le calcul de chaque nombre transformé $X(k)$ à partir des données $x(n)$ est fait avec une erreur $\Delta(k)$ telle que :

$$X(k) + \Delta(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\,[e^{-j2\pi\frac{nk}{N}} + \delta(nk)]$$

soit :

$$\Delta(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\,\delta(nk)$$

Comme il existe entre les $x(n)$ et $X(k)$ la relation (2.2) :

$$x(n) = \sum_{n=0}^{N-1} X(k)\,e^{j2\pi\frac{nk}{N}}$$

il vient :

$$\Delta(k) = \sum_{i=0}^{N-1} X(i)\,\varepsilon(i,k) \qquad (2.18)$$

avec :

$$\varepsilon(i,k) = \frac{1}{N}\sum_{n=0}^{N-1} \delta(nk)\, e^{j2\pi\frac{ni}{N}} \qquad (2.18)$$

Par suite l'arrondi des coefficients de la transformée entraîne pour le nombre transformé $X(k)$ une perturbation $\Delta(k)$ obtenue en faisant une somme de perturbations élémentaires dont chacune est égale au produit d'un nombre transformé par un facteur représentant sa contribution. Les nombres transformés réagissent les uns sur les autres et ne sont plus strictement indépendants.

Pour toute transformée il est possible de calculer les $\varepsilon(i, k)$. Il est en général intéressant de connaître la valeur maximale ε_m que peuvent prendre les $|\varepsilon(i, k)|$ pour un ordre de transformée et un nombre de bits de quantification b_c donnés.

D'après l'inégalité : $|\delta(n)| \leqslant \sqrt{2}.2^{-b_c}$ un maximum pour ε_m est fourni par :

$$\varepsilon_m \leqslant \sqrt{2}.2^{-b_c}$$

En fait les valeurs trouvées en pratique pour ε_m sont nettement inférieures à ce maximum. Par exemple pour $N = 64$, on trouve $\varepsilon_m \simeq 0{,}6.2^{-b_c}$; cette valeur se conserve ensuite pour les valeurs de N supérieures [6].

2.3.2 Bruit de calcul dans la TFR

Les données se présentent à l'entrée d'un calculateur de TFD avec un nombre de bits limité. A chaque opération, addition et multiplication, ce nombre de bits augmente. En général le nombre de bits affecté à chaque donnée reste fixe dans tout le calculateur; il en résulte la nécessité d'opérer des limitations du nombre de bits des nombres au cours du calcul.

Ces limitations sont presque toujours faites par une élimination des bits de plus faible poids avec arrondi. En effet, un dépassement vers les forts poids n'est pas acceptable en général et le cadrage des nombres dans les mémoires doit être étudié en conséquence.

Deux cas importants et simples vont être examinés pour une transformée en base 2. D'abord la transformée directe : avec le facteur d'échelle $\frac{1}{N}$ il suffit de diviser par 2 les résultats des additions et soustractions dans chaque croisillon pour maintenir un cadrage convenable. Ensuite sera étudié le cas où le cadrage au début de la transformée est tel qu'il permet la totalité des calculs sans risque de dépassement. Ce cas peut être celui de la transformée inverse.

Pour évaluer la puissance du bruit de calcul, on considère que la machine stocke les nombres dans des mémoires ayant la capacité de b_i bits pour chaque nombre réel et l'on prend comme unité la plus grande amplitude représentable;

c'est-à-dire que les nombres internes prennent des valeurs comprises dans l'intervalle [– 1, + 1] et que l'échelon de qualification q a pour valeur :

$$q = 2 . 2^{-b_i} = 2^{1-b_i}$$

La figure 2.7 donne le schéma d'un croisillon avec multiplication.

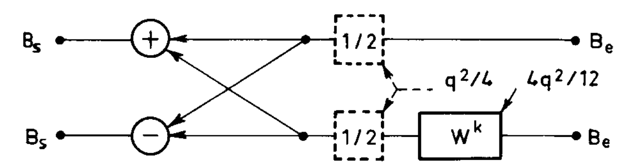


FIG. 2.7. *Croisillon de TFR en base 2*

A l'entrée du croisillon les données complexes sont représentées par une partie réelle et une partie imaginaire comprenant b_i bits chacune. Après multiplication les nombres subissent une limitation à b_i bits avec arrondi. Cette opération sur un nombre réel apporte une puissance de bruit estimée à $\frac{q^2}{12}$. La multiplication complexe se réalisant généralement par 4 multiplications réelles, il y a introduction d'une puissance de bruit égale à $4 \frac{q^2}{12}$. Le coefficient W^k ayant un module qui ne dépasse pas l'unité, aucun recadrage des données dans les mémoires n'est nécessaire.

Bruit de calcul avec recadrage

Le recadrage des données est supposé être réalisé avant les opérations d'addition et soustraction du croisillon, ce qui est un cas défavorable mais facilite la réalisation. Les parties réelle et imaginaire sont divisées par deux par un décalage au cours duquel un bit est éliminé ; ce bit ayant la valeur 0 ou 1 avec la probabilité $\frac{1}{2}$, on admet que sur chaque nombre réel il en résulte une puissance de bruit égale à $\frac{q^2}{8}$ et sur le nombre complexe une puissance de bruit égale à $\frac{q^2}{4}$. Par contre les bruits présents antérieurement sont divisés par 4.

A l'entrée de chaque croisillon le signal est affecté du bruit Be, à la sortie du bruit Bs. Avec le recadrage il existe entre Bs et Be la relation suivante :

$$\text{Bs} = 2\left(\frac{q^2}{4} + \frac{\text{Be}}{4}\right) + \frac{1}{4}\left(4\,\frac{q^2}{12}\right) = \frac{1}{2}\,q^2 + \frac{1}{2}\,\text{Be} + \frac{q^2}{12}$$

Le premier étage de la transformée ne comprenant pas de multiplication ; si le bruit à l'entrée du calculateur de TFR n'est pas pris en considération, il vient en sortie du premier croisillon :

$$\text{Bs}_1 = \frac{q^2}{2}$$

Le second étage, dans l'entrelacement temporel par exemple, a des multiplications par j qui n'entraînent pas d'arrondi.

Le bruit en sortie s'écrit dans ces conditions :

$$\mathrm{Bs}_2 = 2\left(\frac{q^2}{4} + \frac{1}{4}\,\mathrm{Bs}_1\right) = \frac{q^2}{2} + \frac{q^2}{4}$$

de même :

$$\mathrm{Bs}_3 = 2\left(\frac{q^2}{4} + \frac{1}{4}\,\mathrm{Bs}_2\right) + \frac{q^2}{12} = \frac{q^2}{2} + \frac{q^2}{4} + \frac{q^2}{8} + \frac{q^2}{12}$$

Au dernier étage de rang log 2 (N) :

$$\mathrm{B}_{\mathrm{ST}} = \frac{q^2}{2} + \sum_{i=0}^{\mathrm{lb}\left(\frac{\mathrm{N}}{2}\right)} \frac{1}{2^i} + \frac{q^2}{12} \sum_{i=0}^{\mathrm{lb}\left(\frac{\mathrm{N}}{8}\right)} \frac{1}{2^i}$$

Finalement en sortie de transformée on peut écrire :

$$\mathrm{Bs}_{\mathrm{T}} \simeq q^2$$

D'où le résultat : dans une transformée avec recadrage par division par 2 à chaque étage, la puissance de bruit sur chaque sortie peut être estimée à :

$$\mathrm{Bs}_{\mathrm{T}} = 2^{2(1-b_i)} \tag{2.19}$$

Bruit de calcul sans recadrage

Les puissances de bruit à l'entrée et à la sortie d'un croisillon sont liées par la relation :

$$\mathrm{Bs} = 2\,\mathrm{Be} + 4\,\frac{q^2}{12}$$

En considérant qu'il n'y a pas production de bruit dans les 2 premiers étages, le bruit total sur chaque sortie est donné par :

$$\mathrm{B}_{\mathrm{ST}} = 4\,\frac{q^2}{12}\,\frac{\mathrm{N}}{8} \sum_{i=0}^{\mathrm{lb}\left(\frac{\mathrm{N}}{8}\right)} \frac{1}{2^i}$$

d'où :

$$\mathrm{Bs}_{\mathrm{T}} \simeq \mathrm{N}\,\frac{q^2}{12}$$

Dans une transformée sans recadrage, la puissance de bruit sur chaque sortie peut être estimée à :

$$\mathrm{Bs}_{\mathrm{T}} = \mathrm{N}\,.\,\frac{2^{2(1-b_i)}}{12} \tag{2.20}$$

Ce résultat peut s'interpréter en disant que la précision des données se trouve réduite de $\frac{\mathrm{M}}{2}$ bits après le calcul d'une transformée d'ordre $\mathrm{N} = 2^{\mathrm{M}}$.

Le même raisonnement peut être appliqué au calcul suivant d'autres bases, en particulier la base 4. Les résultats obtenus sont comparables.

En pratique les puissances de bruit doivent être associées aux puissances du signal et le paramètre le plus intéressant est le rapport signal à bruit. Pour déterminer comment évolue ce paramètre dans une Transformée, la relation qui lie la puissance du signal à la puissance de son spectre est utilisée :

$$\frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2 = \sum_{k=0}^{N-1}|X(k)|^2$$

Le rapport signal à bruit en sortie de la Transformée dépend de la répartition de la puissance entre les $X(k)$. Par exemple dans le calcul avec recadrage de la Transformée directe, avec l'hypothèse d'une répartition uniforme de la puissance entre les $X(k)$, la puissance de chaque sortie se trouve divisée par N.

Dans ces conditions si S désigne la puissance du signal en entrée, si le bruit à l'entrée est négligé, le rapport signal à bruit en sortie de transformée $(S/B)_{ST}$ s'écrit :

$$(S/B)_{ST} = \frac{S}{N2^{2(1-b_i)}} \qquad (2.21)$$

Dans les mêmes conditions, mais sans recadrage la puissance du signal est multipliée par N et le rapport $(S/B)_{ST}$ devient :

$$(S/B)_{ST} = \frac{S}{2^{2(1-b_i)}/12} \qquad (2.22)$$

Les calculs qui ont été faits dans ce paragraphe doivent être considérés comme approximatifs. Ils ont été menés dans l'hypothèse d'une absence de corrélation entre les erreurs. Une telle hypothèse n'est pas toujours vérifiée, en particulier pour les transformées d'ordre N faible.

L'analyse simplifiée qui a été présentée est suffisante dans la plupart des applications ; une analyse approfondie est donnée dans la référence [7].

L'application la plus directe de la TFD est l'analyse spectrale.

2.4 CALCUL DE SPECTRE PAR TFD

Le calcul d'un spectre par la Transformation de Fourier Discrète oblige à faire certaines approximations et nécessite un choix convenable des paramètres pour atteindre les performances imposées. Avant de considérer les applications il est utile cependant de bien voir la fonction remplie par la Transformation de Fourier Discrète.

2.4.1 Fonction de filtrage remplie par la TFD

Examinons la relation qu'établit la Transformation de Fourier Discrète entre les sorties $X(k)$ et les entrées $x(n)$ considérées comme le résultat de l'échantillonnage d'un signal $x(t)$ avec la période T. Pour $k = 0$ cette relation s'écrit :

$$X(0) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)$$

Le signal X(0) ainsi défini résulte de la convolution du signal $x(t)$ avec la distribution $\varphi_0(t)$ telle que :

$$\varphi_0(t) = \frac{1}{N} \sum_{n=0}^{N-1} \delta(t - nT)$$

La Transformée de Fourier de cette distribution est donnée par :

$$\Phi_0(f) = \frac{1}{N} \sum_{n=0}^{N-1} e^{-j2\pi nfT} = \frac{1}{N} \frac{1 - e^{-j2\pi nfNT}}{1 - e^{-j2\pi nfT}}$$

Ou encore :

$$\Phi_0(f) = e^{-j\pi f(N-1)T} . \Phi(f)$$

avec

$$\Phi(f) = \frac{1}{N} \frac{\sin(\pi fNT)}{\sin(\pi fT)} \tag{2.23}$$

Or une opération de convolution dans l'espace des temps correspond à un produit dans l'espace des fréquences, c'est-à-dire que X(0) est un signal obtenu par filtrage du signal d'entrée par la fonction $\Phi_0(f)$. La figure 2.8 représente la fonction $\Phi(f)$ et la fonction $\varphi(t)$ dont elle est la transformée de Fourier; la fonction $\Phi(f)$ s'annule aux points de l'axe des fréquences multiples entiers de $\frac{1}{NT}$ sauf aux multiples de $\frac{1}{T}$. Elle est périodique et de période $\frac{1}{T}$ conformément aux lois de l'échantillonnage; il s'agit simplement du spectre d'une impulsion de largeur NT échantillonnée.

A la sortie X(k) correspond la fonction $\varphi_k(t)$ telle que :

$$\varphi_k(t) = \frac{1}{N} \sum_{n=0}^{N-1} e^{-j2\pi\frac{nk}{N}} \delta(t - nT)$$

$$\Phi_k(f) = \frac{1}{N} \sum_{n=0}^{N-1} e^{-j2\pi\frac{nk}{N}} e^{-j2\pi fnT}$$

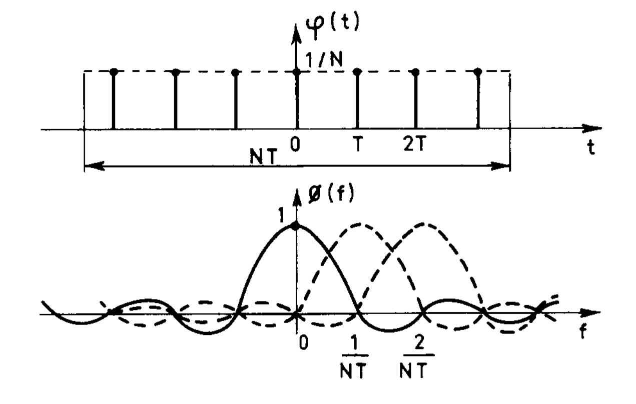


FIG. 2.8. *Fonction de filtrage de la TFD*

Sous forme concise, après simplification, il vient :

$$\Phi_k(f) = (-1)^k\, e^{-j\pi f(N-1)T}\, e^{j\pi \frac{k}{N}}\, \Phi\left(f + \frac{k}{NT}\right) \qquad (2.23\,bis)$$

La sortie $X(k)$ fournit le signal filtré suivant la fonction $\Phi(f)$ translatée de $\frac{k}{NT}$ sur l'axe des fréquences.

Finalement la transformée de Fourier Discrète constitue un ensemble de N filtres identiques, ou banc de filtres, répartis uniformément dans le domaine des fréquences avec le pas $\frac{1}{NT}$.

Si le signal d'entrée est une suite périodique, suivant l'hypothèse de définition de la TFD, ce banc de filtres se trouve échantillonné en fréquence avec le pas $\frac{k}{NT}$, et l'on peut remarquer qu'il n'y a pas d'interférences entre les sorties $X(k)$. Cette propriété cependant est perdue, en toute rigueur, si les coefficients sont arrondis, comme un paragraphe précédent l'a montré.

La mise en évidence de la fonction remplie par la TFD illustre aussi le problème du cadrage des nombres dans les mémoires d'un calculateur de TFR. En effet supposons que les nombres à transformer $x(n)$ résultent de l'échantillonnage d'un signal aléatoire dont la loi de probabilité des amplitudes a la variance σ^2. Si ce signal a une distribution spectrale énergétique uniforme, sa puissance se répartit uniformément entre les $X(k)$, et chacun a une variance égale à $\frac{\sigma^2}{N}$. Par contre si le signal a une distribution spectrale qui peut se concentrer sur un $X(k)$, cet $X(k)$ a la même loi de probabilité que les $x(n)$, en particulier la variance σ^2. Le recadrage des nombres par division par 2 à chaque étage d'un calcul de TFR est adapté au traitement de tels signaux.

On peut approfondir l'étude du processus de filtrage en observant que les sorties $X(k)$ de la TFD sont les sommes des entrées $x(n)$ après déphasage. En effet la sortie $X(0)$ est la somme des $x(n)$ avec déphasages nuls, la sortie $X(k)$ est la somme des $x(n)$ avec déphasages multiples de $2\pi\,\frac{k}{N}$, comme le montre la figure 2.9. À chaque sortie les composantes des signaux résultants qui sont en phase s'ajoutent, les autres s'annulent. Par exemple si les $x(n)$ sont des nombres complexes ayant la même phase et le même module, tous les $X(k)$ s'annulent sauf $X(0)$.

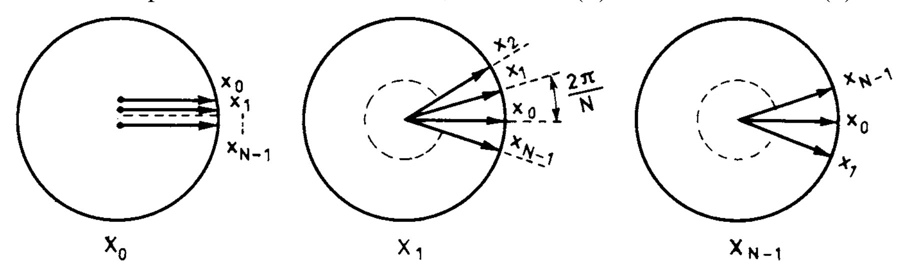


Fig. 2.9. *Filtrage par déphasage dans la TFD*

Cette observation est utile dans l'étude des bancs de filtres qui comprennent un calculateur de TFD.

2.4.2. Résolution spectrale

L'analyse spectrale est utilisée dans de nombreux domaines. Elle se fait à partir d'un enregistrement fourni par un capteur. Or, par définition la TFD établit une relation entre deux suites périodiques, les $x(n)$ et $X(k)$ qui comprennent chacune N éléments différents. Pour l'utiliser il faut donc introduire cette double périodicité.

La périodicité en fréquence est introduite par l'opération d'échantillonnage du signal. L'enregistrement à traiter se présente soit sous forme numérique, alors l'échantillonnage et le codage ont été effectués par le capteur, soit sous forme analogique et alors il faut le numériser. Le choix de la fréquence d'échantillonnage $f_e = \frac{1}{T}$ doit être tel que les composantes du signal de fréquence supérieure à $f_e/2$ soient négligeables et en tout cas inférieures en amplitude à l'erreur tolérée sur l'amplitude des composantes utiles. On peut s'assurer que cette condition est bien remplie en procédant à un préfiltrage du signal.

La périodicité temporelle est introduite artificiellement en supposant que le signal se reproduit en dehors de l'intervalle de temps $\theta = NT$ qui correspond à l'enregistrement à traiter. Dans ces conditions la TFD fournit un échantillonnage du spectre avec une période fréquentielle Δf égale à l'inverse de la durée de l'enregistrement et qui constitue la résolution fréquentielle de l'analyse. La relation $\Delta f = \frac{1}{NT}$ exprime pour l'analyse spectrale la relation d'incertitude de Heisenberg. Une analyse plus fine peut être obtenue en augmentant la durée de l'enregistrement, par exemple en la portant à $N'T$ (avec $N' > N$) avec des échantillons complémentaires nuls ; les échantillons fréquentiels supplémentaires obtenus constituent simplement une interpolation des précédents ; cette opération se pratique couramment pour avoir un nombre de données N' qui soit une puissance de 2 et pour pouvoir utiliser les algorithmes de calcul rapides. D'autre part le fait que le signal ne soit pas composé uniquement de raies aux fréquences multiples de $\frac{1}{NT}$ entraîne une interférence entre les composantes spectrales obtenues ; en effet la fonction de filtrage $\Phi(f)$ de la TFD, qui a été donnée au paragraphe 2.4.1, présente des ondulations dans toute la bande de fréquences et si le signal possède une composante spectrale $S(f_0)$ à la fréquence f_0, c'est-à-dire si $x(t) = S(f_0)\,e^{j2\pi f_0 t}$, on obtient :

$$X(k) = S(f_0) \,.\, \Phi\left(\frac{k}{NT} - f_0\right) \qquad (2.24)$$

pour $0 \leqslant k \leqslant N-1$. Si $\frac{k}{NT} < f_0 < \frac{k+1}{NT}$, il en résulte une contribution non seulement sur les sorties $X(k)$ et $X(k+1)$ de la TFD, mais sur toutes les sorties, comme le montre la figure 2.10 ; ainsi apparaissent les limitations du pouvoir séparateur

de l'analyseur. Cet effet peut être atténué en modifiant la fonction de filtrage de la TFD par pondération des échantillons du signal avant transformation.

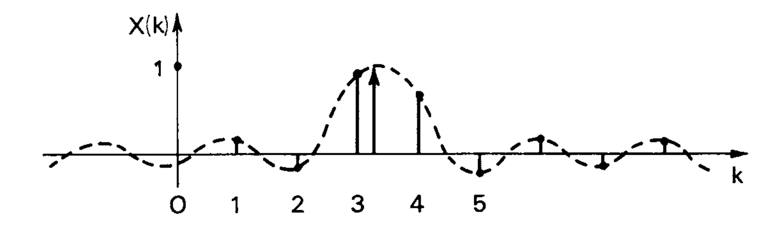


FIG. 2.10. *Analyse d'un signal de fréquence non multiple de* $\frac{1}{NT}$

Cette opération revient à remplacer la fenêtre temporelle rectangulaire $\varphi(t)$, par une fonction dont la transformée de Fourier présente des ondulations plus faibles. De nombreuses fonctions sont utilisées, les plus simples sont la cosinusoïde surélevée :

$$\varphi(t) = \frac{1}{2}\left(1 + \cos 2\pi \frac{t}{\mathrm{NT}}\right) \tag{2.25}$$

et la fenêtre dite de Hamming :

$$\varphi(t) = 0{,}54 + 0{,}46 \cos\left(2\pi \frac{t}{\mathrm{NT}}\right) \tag{2.26}$$

Cette dernière fonction a 99,96 % de son énergie dans le lobe principal et le lobe secondaire le plus important se trouve à environ 40 dB au-dessous du lobe principal. La référence [8] présente un ensemble de fonctions ayant des propriétés variées pour application à l'analyse spectrale.

Si $\Phi(f)$ désigne le spectre de la fonction fenêtre $\varphi(t)$ après échantillonnage, la formule (2.24) se généralise à un signal ayant un spectre quelconque $\mathrm{S}(f)$, en utilisant la définition du produit de convolution et en tenant compter de la périodicité de $\Phi(f)$; il vient :

$$\mathrm{X}(k) = \int_0^{\frac{1}{\mathrm{T}}} \left[\frac{1}{\mathrm{T}} \sum_{n=-\infty}^{\infty} \mathrm{S}\left(u - \frac{n}{\mathrm{T}}\right)\right] \Phi\left(\frac{k}{\mathrm{NT}} - u\right) du \tag{2.27}$$

Cette expression fait apparaître le repliement de bande du signal dû à l'échantillonnage avec la période T.

Pour mieux maîtriser les interférences entre les composantes spectrales calculées, il faut faire appel à un banc de filtres plus sélectifs, comme celui qui est présenté au chapitre 10.

La TFD peut être utilisée d'une manière indirecte dans la procédure de calcul des convolutions.

2.5 LA CONVOLUTION RAPIDE

La puissance de calcul des algorithmes de Transformation de Fourier Rapide conduit à utiliser la TFD dans des cas autres que l'analyse spectrale et en particulier pour la réalisation d'opérations de convolution. Bien que cette approche ne soit pas en général la plus efficace, elle peut présenter de l'intérêt dans les applications où un calculateur de TFR est disponible.

Parmi les propriétés de la TFD figure la propriété suivante : la transformée d'un produit de convolution est égale au produit des transformées. Étant donné deux suites $x(n)$ et $h(n)$ de période N et dont les transformées sont $X(k)$ et $H(k)$, la convolution circulaire :

$$y(n) = \sum_{m=0}^{N-1} h(m)x(n-m)$$

est une suite de même période dont la transformée s'écrit :

$$Y(k) = H(k).X(k).$$

La convolution rapide consiste à calculer la suite $y(n)$ en effectuant une transformée de Fourier Discrète inverse sur la suite $Y(k)$. Comme en général un des termes de la convolution est constant l'opération demande une TFD, un produit et une TFD inverse. Cette technique s'applique aux suites de durée finie ; si $x(n)$ et $h(n)$ sont deux suites de N1 et N2 termes non nuls, la suite $y(n)$ définie par :

$$y(n) = \sum_{m=0}^{n} h(m)x(n-m)$$

est une suite de durée finie, ayant $N_1 + N_2 - 1$ termes ; la convolution rapide s'applique en considérant que les trois suites $y(n)$, $x(n)$ et $h(n)$ ont la période N tel que $N \geqslant N_1 + N_2 - 1$; il suffit de compléter par un nombre convenable de termes nuls. Il est alors intéressant de choisir pour N une puissance de deux.

Cependant en pratique l'opération de convolution est une opération de filtrage, où les $x(n)$ représentent le signal et les $h(n)$ les coefficients. La suite des $x(n)$ est beaucoup plus longue que la suite des $h(n)$ et il faut fractionner le calcul. Dans ce but la suite des $x(n)$ est considérée comme une superposition de suites élémentaires $x_k(n)$ de N_3 termes. C'est-à-dire que l'on a :

$$x(n) = \sum_k x_k(n)$$

avec $x_k(n) = x(n)$ pour $kN_3 \leqslant n \leqslant (k+1)N_3 - 1$ et $x_k(n) = 0$ ailleurs.

Alors on peut écrire :

$$y(n) = \sum_{m=0}^{n} h(m) \sum_k x_k(n-m)$$

$$y(n) = \sum_k \sum_{m=0}^{n} h(m)\, x_k(n-m) = \sum_k y_k(n)$$

Chaque suite $y_k(n)$ compte $N_3 + N_2 - 1$ termes non nuls. Les convolutions à réaliser portent sur $N_3 + N_2 - 1$ termes. Le diagramme de la figure 2.11 montre l'enchaînement des opérations ; les suites $y_k(n)$ et $y_{k+1}(n)$ ont $N_2 - 1$ termes qui se superposent.

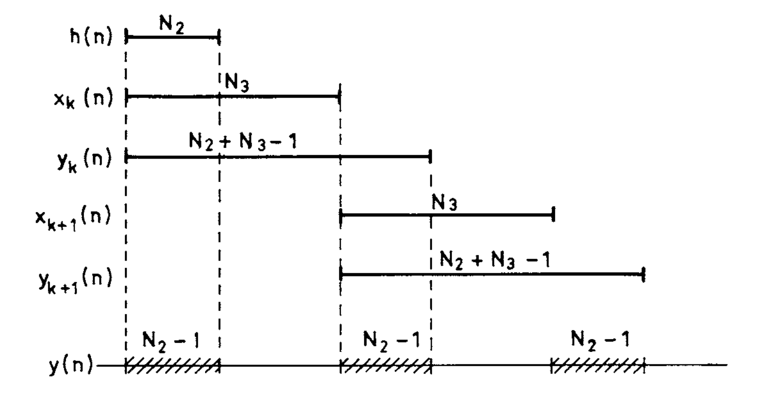


FIG. 2.11. *Enchaînement des opérations dans la convolution rapide avec fractionnement*

Dans ce processus le nombre de calculs à effectuer par élément de la suite $y(n)$ croît comme $\mathrm{Log}_2(N_2 + N_3 - 1)$ et il ne faut pas choisir N_3 trop grand ; si $N_3 < N_2$ aucun terme de la suite $y(n)$ n'est obtenu directement. Par suite il existe une valeur optimale pour N_3. Le nombre de positions de mémoires nécessaire croît comme $N_3 + N_2 + 1$. Un bon compromis consiste à prendre pour N_3 la première valeur supérieure à N_2 telle que $N_3 + N_2 - 1$ soit une puissance de deux.

2.6 CALCUL D'UNE TFD PAR CONVOLUTION

Dans certaines applications on ne dispose que d'opérateurs capables de faire des convolutions pour calculer une TFD. C'est le cas des circuits utilisant les dispositifs à transfert de charges qui permettent de faire des calculs, sous forme analogique et sur des signaux échantillonnés, à des vitesses compatibles avec les fréquences rencontrées dans le domaine des radars par exemple.

La relation de définition de la TFD s'écrit :

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-j2\pi \frac{nk}{N}}$$

En posant

$$nk = \frac{1}{2} [n^2 + k^2 - (n-k)^2] \quad \text{et} \quad W = e^{-j\frac{2\pi}{N}}$$

il vient :

$$X(k) = W^{\frac{k^2}{2}} \sum_{n=0}^{N-1} x(n)\, W^{\frac{\pi^2}{2}}\, W^{\frac{-(n-k)^2}{2}} \qquad (2.28)$$

Cette expression exprime le produit de convolution circulaire des suites $x(n)W^{\frac{n^2}{2}}$ et $W^{-\frac{n^2}{2}}$. Il s'en suit que le calcul des $X(k)$ peut être effectué en trois étapes comprenant les opérations suivantes :

– multiplication des données $x(n)$ par les coefficients $W^{\frac{n^2}{2}}$

– produit de convolution par la suite de coefficients $W^{-\frac{n^2}{2}}$

– multiplication des résultats par les coefficients $W^{\frac{k^2}{2}}$

Le processus est représenté sur la figure 2.12.

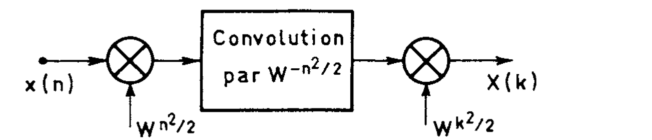


FIG. 2.12. *Calcul d'une TFD par convolution*

Cette méthode s'étend au cas où W est un nombre complexe dont le module est différent de l'unité [9].

Après la présentation et l'application des algorithmes de TFR, certains aspects de la réalisation vont être abordés.

2.7 RÉALISATION

Pour mettre en œuvre les algorithmes de calcul de la TFD il faut disposer d'équipements comprenant les éléments suivants :

– unité de mémoire pour stocker les données d'entrée-sortie et les résultats intermédiaires;

– unité de mémoire pour stocker les coefficients de la transformée;

– unité arithmétique capable d'effectuer des additions et multiplications portant sur des nombres complexes;

– unité de commande pour enchaîner les opérations.

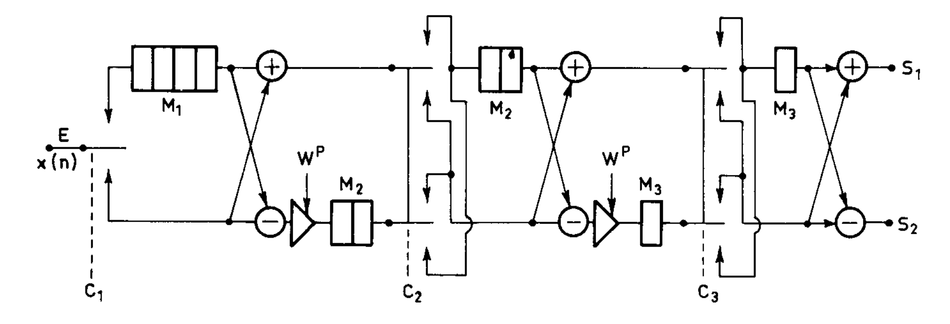
Ces éléments de base se retrouvent dans toute machine destinée au traitement numérique du signal, que ce soit en logique câblée ou en logique programmée. Les particularités liées à la mise en œuvre de la TFR tiennent à deux caractéristiques principales :

– un volume de calculs arithmétiques important;

– les permutations à effectuer sur les données qui conduisent à des calculs d'indices compliqués.

Les contraintes sont d'autant plus grandes que l'ordre de la Transformée est élevé.

Le problème de réalisation qui se pose est celui d'une mise en œuvre efficace des algorithmes décrits dans les paragraphes précédents. Une mise en œuvre efficace est celle qui conduit à un taux élevé d'utilisation des différentes unités de la machine et particulièrement l'unité arithmétique. On peut imaginer différents agencements de circuits et modes d'organisation des calculs [10] et [11]. A titre d'exemple, un circuit simple pour réaliser l'algorithme de la figure 2.4 est donné par la figure 2.13. Les données $x(n)$ se présentent en série, la Transformée est d'ordre N = 8.



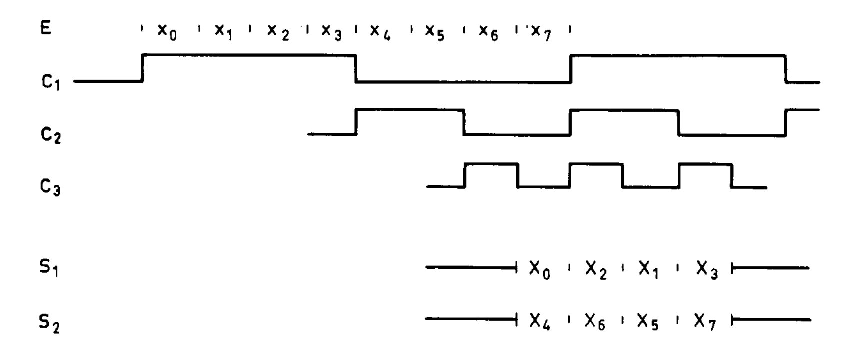


FIG. 2.13. *Réalisation série d'une TFR en base 2*

Le circuit comporte un registre M_1 de 4 données, deux registres M_2 de 2 données et 2 registres M_3 d'une donnée. Les dispositifs de commutation, commandés par les signaux C1, C2 et C3 respectivement permettent un enchaînement convenable des calculs. Les coefficients W^P doivent être appliqués aux multiplieurs au moment approprié. La figure donne les signaux de commande et les signaux d'entrée et sortie.

Le multiplieur complexe est réalisé à l'aide d'un ensemble de multiplieurs et additionneurs réels qui pour faire l'opération :

$$C_R + jC_i = (a_R + ja_i)(b_R + jb_i)$$

effectuent les calculs suivants :

$$C_R = a_R b_R - a_i b_i$$
$$C_i = a_R b_i + a_i b_R$$

Il faut noter qu'il existe une autre possibilité qui conduit pour chaque multiplication complexe à 3 multiplications réelles au lieu de 4 et 5 additions au lieu de 2 :

$$(a_R + ja_i)(b_R + jb_i) = [(a_R + a_i)\,b_R - a_i(b_R + b_i)] + j[(a_R + a_i)\,b_R + a_R(b_i - b_R)] \quad (2.29)$$

Et même, si le nombre complexe $b_R + jb_i$ est fixe, le nombre d'additions à faire se réduit à 3 pour chaque multiplication.

Si le débit de données est continu et que les calculs de TFR sont à faire sur des blocs de données adjacents, le rythme de base des calculs est celui des données et l'on peut remarquer que les multiplieurs sont inactifs pendant la moitié du temps. Il est possible de perfectionner le circuit pour que les multiplieurs soient utilisés à leur capacité maximale, en traitant des blocs de données consécutifs avec un rythme de base égal à la moitié de la cadence des données.

L'exemple de réalisation décrit ci-dessus appelle deux remarques :

– Le circuit comprend trois étages de structure différente, avec des signaux de commande différents ; c'est un inconvénient car on recherche plutôt les circuits modulaires, en particulier pour l'intégration à grande échelle.

– Les résultats se présentent dans un ordre permuté ; il faut un traitement complémentaire pour retrouver l'ordre naturel des indices.

Quand un certain degré d'optimisation est recherché, plutôt qu'adapter les circuits aux algorithmes, il est préférable de rechercher des algorithmes compatibles avec les contraintes qu'impose la technologie aux circuits.

D'autre part, en vue de réduire encore le volume des calculs arithmétiques, des algorithmes plus sophistiqués que la TFR peuvent être envisagés. De plus si les données présentent des particularités, par exemple si ce sont des nombres réels, si des symétries apparaissent, des simplifications importantes peuvent intervenir.

Ces sujets sont abordés dans le chapitre suivant, à partir d'une représentation unifiée des algorithmes de TFR.

BIBLIOGRAPHIE

[1] Special Issue on FFT and applications. *IEEE Transactions,* Vol AU-15, N° 2, June 1967.

[2] A. Oppenheim and R. Schafer – *Digital Signal Processing* chapters 3 and 6, Prentice-Hall, New Jersey, 1974.

[3] L. Rabiner and B. Gold – *Theory and Application of Digital Signal Processing.* Chapters 6 and 10, Prentice-Hall, New Jersey, 1975.

[4] J. Liferman – *Théorie et Application de la Transformation de Fourier Rapide.* Éd. Masson, Paris, 1977.

[5] P. Duhamel – « Un algorithme de Transformation de Fourier Rapide à double base », Annales des Télécom., Tome 40, N° 9-10, Sept.-Oct. 1985, pp. 481-494.

[6] D. Tufts and H. Hersey – *Effects of FFT Coefficient Quantization on Bin Frequency Response.* Proceedings of IEEE, January 1972.

[7] Tran-Thong and Bede Liu – Fixed Point FFT Error Analysis. *IEEE Trans.* ASSP, Vol. 24, N° 6, December 1976.

[8] J. Max et Collaborateurs – *Méthodes et Techniques de Traitement du Signal et Applications aux Mesures Physiques,* Tome I, Éditions Masson, Paris, 1981.

[9] L. R. Rabiner, R. W. Schafer and C. M. Rader – The Chirp Z-Transform Algorithm. *IEEE Trans.,* Vol. AU-17, June 1969.

[10] H. L. Groginsky and G. A. Works – A Pipeline Past Fourier Transform, *IEEE Trans.* On Computers, C 19, Nov. 1970.

[11] B. Gold and T. Bially – Parallelism in Fast Fourier Transform Hardware, *IEEE Trans.* AU-21, N° 1, Feb. 1973.

EXERCICES

1 Calculer la Transformée de Fourier Discrète de la suite comportant N = 16 termes tels que :

$$x(0) = x(1) = x(2) = x(14) = x(15) = 1$$
$$x(n) = 0 \quad \text{pour} \quad 3 \leqslant n \leqslant 13$$

et de la suite :

$$x(0) = x(1) = x(2) = x(3) = x(4) = 1$$
$$x(n) = 0 \quad \text{pour} \quad 5 \leqslant n \leqslant 15$$

Comparer les résultats obtenus. Effectuer la Transformée inverse sur ces résultats.

2 Établir le diagramme de l'algorithme de TFR d'ordre 16 avec entrelacement temporel et entrelacement fréquentiel. Quel est le nombre minimum de multiplications et additions nécessaires ?

3 Calculer, en utilisant le programme donné en annexe, la Transformée de Fourier Discrète de la suite comportant N = 128 termes tels que :

$$x(0) = x(1) = x(2) = x(126) = x(127) = 1$$
$$x(n) = 0 \quad \text{pour} \quad 3 \leqslant n \leqslant 125$$

Comparer les résultats à ceux donnés par l'exercice 1.

La suite $X(k)$ obtenue constitue une approximation du développement en série de Fourier d'une suite d'impulsions. Rapprocher les résultats obtenus avec les chiffres du tableau de l'annexe I du chapitre I. Expliquer les différences. Comment évolue l'approximation si N = 512 ?

4 Soit à réaliser une TFD d'ordre 64 avec le minimum d'opérations arithmétiques.

Évaluer le nombre de multiplications et d'additions nécessaires avec les algorithmes en base 2, 4 et 8.

5 Soit à analyser la puissance du bruit de calcul produite dans une Transformée d'ordre 32. Montrer, en utilisant les résultats du paragraphe 2.3, comment varient les résultats sur les différentes sorties. Calculer la distorsion introduite par une limitation à 8 bits des coefficients.

6 Montrer que chaque sortie d'une TFD, $X(k)$, peut être obtenue à partir des entrées $x(n)$ par une relation de récurrence. Évaluer le nombre de multiplications nécessaire.

7 Soit à réaliser une TFD d'ordre 64 sur des données qui sont des nombres de 16 bits. Évaluer la dégradation du rapport signal à bruit quand on fait en cascade une Transformée directe et inverse, avec une machine à 16 bits.

8 La bande supposée occupée par un signal à analyser s'étend de 0 à 10 kHz. La résolution spectrale recherchée est de 1 Hz. Quelle longueur d'enregistrement doit-on prélever pour faire une telle analyse ? Quelle capacité de mémoire faut-il pour stocker les données supposées codées à 8 bits ? Déterminer les caractéristiques d'un calculateur capable de réaliser une telle analyse spectrale : capacité de mémoire, cycle mémoire, temps d'addition et multiplication.

9 Calculer la TFD de la suite $x(n)$ définie par :

$$x(n) = \sin 2\pi \frac{n}{3{,}5} + 0{,}2 \sin 2\pi \frac{n}{6{,}5} \quad \text{avec} \quad 0 \leqslant n \leqslant 15.$$

Pour améliorer l'analyse on utilise les fenêtres suivantes :

$$g(n) = 1/2 \left[1 - \cos \frac{2\pi n}{16}\right]$$

$$g(n) = 0{,}54 - 0{,}46 \cos \frac{2\pi n}{16} \quad \text{(Hamming)}$$

$$g(n) = 0{,}42 - 0{,}5 \cos \frac{2\pi n}{16} + 0{,}08 \cos \frac{4\pi n}{16} \quad \text{(Blakman)}$$

Comparer les résultats de cette analyse.

10 Décrire de manière détaillée le fonctionnement du circuit de la figure 2.13. Donner en particulier la suite des nombres qui se présentent en sortie de chaque additionneur et

soustracteur. Montrer que pour faire fonctionner les multiplieurs à leur pleine capacité, il faut introduire une mémoire d'entrée supplémentaire ; donner les signaux de commande dans ce cas.

11 Généraliser à la base 4 la réalisation série donnée au paragraphe 2.7. Donner le diagramme des temps détaillé.

Si la Transformée est d'ordre N = 64, si les données ont 16 bits et se présentent à la cadence de 8 kHz, si les coefficients ont 16 bits, exprimer la puissance de calcul nécessaire et le volume de mémoire pour les données et les coefficients. Rapprocher ces résultats des caractéristiques d'un microprocesseur courant.

Chapitre 3

Autres algorithmes de calcul rapide de la TFR

Les algorithmes de calcul rapide de la Transformée de Fourier Discrète (TFD) sont basés sur une factorisation de la matrice de cette Transformée. Cette factorisation est déjà apparue dans les algorithmes à entrelacement temporel et fréquentiel introduits dans le chapitre précédent et qui sont des éléments d'un ensemble d'algorithmes très variés.

Pour appréhender l'ensemble des algorithmes de calcul rapide et pouvoir ainsi exploiter au mieux les particularités des signaux à traiter et les possibilités technologiques, il est nécessaire de faire appel à un outil mathématique bien adapté, le produit de Kronecker des matrices. En combinant ce produit avec le produit au sens habituel il est possible de décomposer simplement toute matrice de TFD.

3.1 LE PRODUIT DE KRONECKER DES MATRICES

Le produit de Kronecker est une opération du calcul tensoriel qui constitue une généralisation de la multiplication d'une matrice par un scalaire [1].

Étant donné deux matrices A et B ayant respectivement m et p lignes et n et q colonnes, le produit de Kronecker de A par B, noté A $\otimes$ B, est une nouvelle matrice à mp lignes et nq colonnes, que l'on obtient en remplaçant chaque élément b_{ij} de la matrice B par le tableau b_{ij} A suivant :

$$\begin{matrix} b_{ij}a_{11} & b_{ij}a_{12} & \dots & b_{ij}a_{1n} \\ \vdots & & & \vdots \\ b_{ij}a_{m1} & b_{ij}a_{m2} & \dots & b_{ij}a_{mn} \end{matrix}$$

Ce produit n'est généralement pas commutatif :

$$\mathrm{A} \otimes \mathrm{B} \neq \mathrm{B} \otimes \mathrm{A}$$

Par exemple, si la matrice B est telle que

$$B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

Le produit de Kronecker de la matrice A par la matrice B est défini par :

$$A \otimes B = \begin{bmatrix} b_{11}A & b_{12}A \\ b_{21}A & b_{22}A \end{bmatrix} \tag{3.1}$$

On peut remarquer en particulier que le produit de Kronecker de la matrice unité de dimension N, I_N, par la matrice unité de dimension M, I_M est égal à la matrice unité de dimension MN :

$$I_N \otimes I_M = I_{NM} \tag{3.2}$$

De même le produit de Kronecker d'une matrice diagonale par une autre matrice diagonale est encore une matrice diagonale.

Le produit de Kronecker se combine avec le produit de matrices au sens habituel, avec les propriétés suivantes, qui vont être utilisées dans les prochains paragraphes, pourvu que les dimensions soient compatibles :

– Le produit de Kronecker d'un produit de matrices par la matrice unité est égal au produit des produits de Kronecker de chaque matrice par la matrice unité :

$$(ABC) \otimes I = (A \otimes I)(B \otimes I)(C \otimes I) \tag{3.3}$$

– Un produit de produits de Kronecker est égal au produit de Kronecker des produits :

$$(A \otimes B \otimes C)(D \otimes E \otimes F) = (AD) \otimes (BE) \otimes (CF) \tag{3.4}$$

– L'inverse d'un produit de Kronecker est égal au produit de Kronecker des inverses :

$$(A \otimes B \otimes C)^{-1} = A^{-1} \otimes B^{-1} \otimes C^{-1} \tag{3.5}$$

L'opération de transposition a la même propriété que l'inversion :

$$(A \otimes B \otimes C)^t = A^t \otimes B^t \otimes C^t \tag{3.6}$$

La matrice transposée d'un produit de Kronecker est le produit de Kronecker des matrices transposées.

Ces propriétés sont aisément mises en évidence sur des exemples simples; elles sont utilisées pour la factorisation des matrices à éléments redondants et en particulier les matrices de la TFD [2]. L'entrelacement fréquentiel est considéré en premier lieu.

Il faut noter que le facteur d'échelle $\frac{1}{N}$ n'est pas pris en compte dans la suite de ce chapitre.

3.2 FACTORISATION DE LA MATRICE DE L'ALGORITHME D'ENTRELACEMENT FRÉQUENTIEL

Dans les algorithmes examinés au chapitre précédent, l'une des suites, en entrée ou en sortie, se trouve permutée ; la matrice qui représente cet algorithme est une matrice qui se déduit de la matrice T_N par permutation des lignes ou des colonnes suivant que l'on considère l'entrelacement fréquentiel ou temporel [3].

Désignons par T'_N la matrice qui correspond à l'entrelacement fréquentiel et est obtenue par permutation des lignes de la matrice T_N définie par la règle suivante : chaque ligne est numérotée en binaire ; on inverse l'ordre des chiffres de ce nombre binaire ; la valeur du nombre binaire ainsi obtenu indique le numéro de la ligne dans la nouvelle matrice. Par exemple pour N = 8 il vient :

$$T_8 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & W & W^2 & W^3 & -1 & -W & -W^2 & -W^3 \\ 1 & W^2 & -1 & -W^2 & 1 & W^2 & -1 & -W^2 \\ 1 & W^3 & -W^2 & W & -1 & -W^3 & W^2 & -W \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -W & W^2 & -W^3 & -1 & W & -W^2 & W^3 \\ 1 & -W^2 & -1 & W^2 & 1 & -W^2 & -1 & W^2 \\ 1 & -W^3 & -W^2 & -W & -1 & W^3 & W^2 & W \end{bmatrix} \begin{matrix} 000=0 \\ 001=1 \\ 010=2 \\ 011=3 \\ 100=4 \\ 101=5 \\ 110=6 \\ 111=7 \end{matrix}$$

$$T'_8 = \begin{bmatrix} 1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & W^2 & -1 & -W^2 & 1 & W^2 & -1 & -W^2 \\ 1 & -W^2 & -1 & W^2 & 1 & -W^2 & -1 & W^2 \\ 1 & W & W^2 & W^3 & -1 & -W & -W^2 & -W^3 \\ 1 & -W & W^2 & -W^3 & -1 & W & -W^2 & W^3 \\ 1 & W^3 & -W^2 & W & -1 & -W^3 & W^2 & -W \\ 1 & -W^3 & -W^2 & -W & -1 & W^3 & W^2 & W \end{bmatrix} \begin{matrix} 000=0 \\ 100=4 \\ 010=2 \\ 110=6 \\ 001=1 \\ 101=5 \\ 011=3 \\ 111=7 \end{matrix}$$

On peut remarquer que pour N = 2, la matrice T'_2 est égale à T_2, matrice de la Transformée d'ordre 2.

La matrice T'_N se factorise en faisant apparaître la matrice $T'_{\frac{N}{2}}$ et la matrice diagonale $D_{\frac{N}{2}}$ dont les éléments sont les nombres W^k avec $0 \leqslant k \leqslant \frac{N}{2} - 1$. En effet :

$$T'_N = \begin{bmatrix} T'_{\frac{N}{2}} & T'_{\frac{N}{2}} \\ T'_{\frac{N}{2}} D_{\frac{N}{2}} & -T'_{\frac{N}{2}} D_{\frac{N}{2}} \end{bmatrix}$$

Cette décomposition apparaît clairement pour T'_8 par exemple. Si $I_{\frac{N}{2}}$ désigne la matrice unité d'ordre $\frac{N}{2}$ il vient :

$$T'_N = \begin{bmatrix} T'_{\frac{N}{2}} & 0 \\ 0 & T'_{\frac{N}{2}} \end{bmatrix} \begin{bmatrix} I_{\frac{N}{2}} & I_{\frac{N}{2}} \\ D_{\frac{N}{2}} & -D_{\frac{N}{2}} \end{bmatrix}$$

ou encore :

$$T'_N = \begin{bmatrix} T'_{\frac{N}{2}} & 0 \\ 0 & T'_{\frac{N}{2}} \end{bmatrix} \begin{bmatrix} I_{\frac{N}{2}} & 0 \\ 0 & D_{\frac{N}{2}} \end{bmatrix} \begin{bmatrix} I_{\frac{N}{2}} & I_{\frac{N}{2}} \\ I_{\frac{N}{2}} & -I_{\frac{N}{2}} \end{bmatrix}$$

En utilisant les produits de Kronecker des matrices, on obtient pour T'_N :

$$T'_N = (T'_{\frac{N}{2}} \otimes I_2)\, \Delta_N\, (I_{\frac{N}{2}} \otimes T'_2) \qquad (3.7)$$

où Δ_N est une matrice carrée d'ordre N diagonale, dont les $\frac{N}{2}$ premiers éléments valent 1 et les suivants sont les puissances de W, W^k avec $0 \leqslant k \leqslant \frac{N}{2} - 1$.

Par itération on obtient la factorisation complète :

$$T'_N = (T'_2 \otimes I_{\frac{N}{2}})\,(\Delta_4 \otimes I_{\frac{N}{4}})(I_2 \otimes T'_2 \otimes I_{\frac{N}{4}})$$

....

$$(\Delta_{\frac{N}{2}} \otimes I_2)(I_{\frac{N}{4}} \otimes T'_2 \otimes I_2)$$

$$\Delta_N(I_{\frac{N}{2}} \otimes T'_2)$$

ou encore :

$$T'_N = \sum_{i=1}^{\mathrm{Log}_2 N} (\Delta_{2^i} \otimes I_{\frac{N}{2^i}})(I_{2^{i-1}} \otimes T'_2 \otimes I_{\frac{N}{2^i}}) \qquad (3.8)$$

Cette expression montre que la transformée se calcule en $\mathrm{Log}_2(N)$ étapes qui chacune comprennent :

– une partie d'arrangement des données correspondant au facteur $(I_{2^{i-1}} \otimes T'_2 \otimes I_{\frac{N}{2^i}})$ et qui ne comprend que des additions et soustractions,

– une partie comprenant les multiplications par les coefficients représentés dans la matrice $(\Delta_{2^i} \otimes I_{\frac{N}{2^i}})$.

L'étape correspondant à $i = 1$ ne comprend pas de multiplications. On peut vérifier que toutes les matrices ont bien la dimension N.

Pour voir comment la factorisation se généralise à la base 4, il est intéressant d'examiner la matrice T'_{16}, obtenue à partir de la matrice T_{16} par la permutation suivante des lignes : les lignes sont numérotées en base 4 ; on inverse l'ordre des chiffres dans les numéros de ligne ; la valeur obtenue indique le numéro de la ligne dans la nouvelle matrice. Suivant cette permutation on a : $T_4 = T'_4$.

Avec les notations suivantes :

$$D_4^i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & W^i & 0 & 0 \\ 0 & 0 & W^{2i} & 0 \\ 0 & 0 & 0 & W^{3i} \end{bmatrix}$$

la matrice de la transformée d'ordre 16 ainsi obtenue s'écrit :

$$T'_{16} = \begin{bmatrix} T_4 & T_4 & T_4 & T_4 \\ T_4 D_4^1 & T_4(-j)D_4^1 & T_4(-1)D_4^1 & T_4(+j)D_4^1 \\ T_4 D_4^2 & T_4(-1)D_4^2 & T_4 D_4^2 & T_4(-1)D_4^2 \\ T_4 D_4^3 & T_4(+j)D_4^3 & T_4(-1)D_4^3 & T_4(-j)D_4^3 \end{bmatrix}$$

$$T'_{16} = \begin{bmatrix} T_4 & 0 & 0 & 0 \\ 0 & T_4 & 0 & 0 \\ 0 & 0 & T_4 & 0 \\ 0 & 0 & 0 & T_4 \end{bmatrix} \begin{bmatrix} I_4 & 0 & 0 & 0 \\ 0 & D_4^1 & 0 & 0 \\ 0 & 0 & D_4^2 & 0 \\ 0 & 0 & 0 & D_4^3 \end{bmatrix}$$

$$\begin{bmatrix} I_4 & I_4 & I_4 & I_4 \\ I_4 & -jI_4 & -I_4 & +jI_4 \\ I_4 & -I_4 & I_4 & -I_4 \\ I_4 & +jI_4 & -I_4 & -jI_4 \end{bmatrix}$$

Sous forme de produits de Kronecker cette expression s'écrit :

$$T'_{16} = (T_4 \otimes I_4)\Delta_{16}(I_4 \otimes T_4) \tag{3.9}$$

Δ_{16} est une matrice diagonale dont les 4 premiers termes ont pour valeur 1, les 4 suivants W^k avec $0 \leqslant k \leqslant 3$, les suivants $(W^2)^k$ et $(W^3)^k$ avec $0 \leqslant k \leqslant 3$.

La factorisation en produits de Kronecker est à la base d'algorithmes ayant des propriétés variées, notamment sur l'ordre de présentation et d'extraction des données et l'enchaînement des opérations.

Elle s'applique aussi aux transformées partielles qui ont une grande importance pratique.

3.3 LES TRANSFORMÉES PARTIELLES

Les transformées qui ont été étudiées dans les paragraphes précédents portent sur des ensembles de N nombres qui peuvent être complexes. Dans une analyse

spectrale fine il se peut que l'ordre de la transformée N devienne très grand alors que seuls sont intéressants à connaître un nombre réduit de points du spectre. La limitation du calcul aux seuls points utiles peut alors permettre une économie importante.

Soit à calculer la transformée partielle définie par l'équation suivante où r est un diviseur de N :

$$\begin{bmatrix} X_p \\ X_{p+1} \\ X_{p+2} \\ \vdots \\ X_{p+r-1} \end{bmatrix} = \begin{bmatrix} 1 & W^{P} & W^{2P} & \dots & W^{(N-1)P} \\ 1 & W^{P+1} & W^{2(P+1)} & \dots & W^{(N-1)(P+1)} \\ 1 & W^{P+2} & W^{2(P+2)} & \dots & W^{(N-1)(P+2)} \\ \vdots & \vdots & & \dots & \vdots \\ 1 & W^{(P+r-1)} & \dots\dots\dots & \dots & W^{(N-1)(P+r-1)} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{N-1} \end{bmatrix} \quad (3.10)$$

A partir de l'ensemble des données, on forme $\frac{N}{r}$ sous ensembles de r termes chacun :

$$(x_0, x_{\frac{N}{r}}, x_{\frac{2N}{r}}, \quad , \quad x_{\frac{(r-1)N}{r}})$$

$$(x_1, x_{\frac{N}{r}+1}, \quad , \quad x_{\frac{(r-1)N}{r}+1})$$

$$\dots\dots\dots\dots\dots \quad \dots\dots\dots\dots\dots$$

$$(x_{\frac{N}{r}-1}, x_{\frac{2N}{r}-1}, \quad , \quad x_{N-1})$$

Soit D_r^i la matrice diagonale de dimension r et dont les éléments de la diagonale sont les puissances de W^i, soit W^{ik}, avec $0 \leqslant k \leqslant r-1$.

La matrice de la transformée partielle peut être séparée en $\frac{N}{r}$ sous-matrices qui s'appliquent chacune à l'une des suites définies précédemment et l'équation matricielle de la transformée s'écrit :

$$[X]_{p,r} = \sum_{i=0}^{\frac{N}{r}-1} D_r^i T_r (W^p)^i D_r^{p\frac{N}{r}} [x]_{i,r} \quad (3.11)$$

où $[X]_{p,r}$ désigne l'ensemble de r nombres X_k avec $p \leqslant k \leqslant p+r-1$ et $[x]_{i,r}$ l'ensemble des données x_k avec $k = n \cdot \frac{N}{r} + i$ et $n = 0, 1, \dots r-1$. D'autre part T_r est la transformée d'ordre r.

Par suite si r est un diviseur de N une transformée partielle portant sur r points se calcule à l'aide de $\frac{N}{r}$ transformées d'ordre r auxquelles sont associées des matrices diagonales convenables.

Si N et r sont des puissances de 2, le nombre de multiplications complexes à effectuer M_P s'exprime par :

$$M_P = \frac{N}{r}\left(\frac{r}{2}\log 2\left(\frac{r}{2}\right) + 2r\right) = N\left[\frac{1}{2}\log 2\left(\frac{r}{2}\right) + 2\right] \qquad (3.12)$$

Par rapport à la transformée globale, il apparaît ainsi que la transformée partielle est avantageuse pour $N > 16r$, donc pour les faibles nombres de points à calculer.

Ce résultat est également valable lorsque c'est le nombre de points à transformer qui est limité, ce qui est aussi un cas fréquent en analyse de spectre.

Un exemple courant de transformée partielle est celle qui porte sur des données réelles.

3.3.1 Transformée de nombres réels et TFD impaire

Si les nombres à transformer sont réels, les propriétés énoncées au chapitre 2 indiquent que les nombres transformés $X(k)$ et $X(N - k)$ sont complexes conjugués, c'est-à-dire que $X(k) = \overline{X}(N - k)$; alors il suffit de calculer l'ensemble des X_k avec $0 \leqslant k \leqslant \frac{N}{2} - 1$ et le résultat précédent s'applique :

$$[X]_{0,\frac{N}{2}} = T_{\frac{N}{2}}[x]_{0,\frac{N}{2}} + D_{\frac{N}{2}}T_{\frac{N}{2}}[x]_{1,\frac{N}{2}} \qquad (3.13)$$

Dans ce cas précis il est possible de n'effectuer qu'une seule fois le calcul de la transformée $T_{\frac{N}{2}}$, en tenant compte de la propriété suivante de la Transformée de Fourier Discrète; si la suite à transformer x_k est purement imaginaire, la suite transformée est telle que :

$$X(k) = -\overline{X}(N - k)$$

Dans ces conditions la procédure pour le calcul de la transformée d'une suite réelle est la suivante :

– Former à partir des $x(k)$ une suite complexe de $\frac{N}{2}$ termes

$y(k) = x(2k) + jx(2k + 1)$ avec $0 \leqslant k \leqslant \frac{N}{2} - 1$.

– Calculer la transformée $Y(k)$ de la suite $y(k)$ avec $0 \leqslant k \leqslant \frac{N}{2} - 1$.

– Calculer les nombres cherchés par l'expression :

$$X(k) = \frac{1}{2}\left[Y(k) + \overline{Y}\left(\frac{N}{2} - k\right)\right] + \frac{1}{2}je^{-j2\pi\frac{k}{N}}\left[\overline{Y}\left(\frac{N}{2} - k\right) - Y(k)\right] \qquad (3.14)$$

avec $0 \leqslant k \leqslant \frac{N}{2}$ et $Y\left(\frac{N}{2}\right) = Y(0)$.

En regroupant les facteurs on peut écrire sous une autre forme :

$$X(k) = A(k)\,Y(k) + B(k)\,\overline{Y}\left(\frac{N}{2} - k\right) \tag{3.15}$$

avec $A(k) = \frac{1}{2}(1 - jW^k)$ et $B(k) = \frac{1}{2}(1 + jW^k)$.

La transformée inverse s'obtient à partir des N+1 termes $X(k)$ avec $0 \leqslant k \leqslant \frac{N}{2}$, en calculant :

$$Y(k) = \overline{A}(k)\,X(k) + \overline{B}(k)\,\overline{X}\left(\frac{N}{2} - k\right) \tag{3.16}$$

pour $0 \leqslant k \leqslant \frac{N}{2} - 1$, puis en calculant la TFD inverse d'ordre $\frac{N}{2}$ de ces valeurs et enfin en prenant les parties réelles de l'ensemble obtenu comme données d'indices impairs et les parties imaginaires commes données d'indices pairs.

Si N est une puissance de 2, le nombre de multiplications complexes Mc à effectuer s'élève à :

$$Mc = \frac{N}{4} \log 2\left(\frac{N}{4}\right) + \frac{N}{2} = \frac{N}{4} \log 2\,(N) \tag{3.17}$$

Le nombre de mémoires nécessaire est de N positions réelles. La référence [4] décrit en détail un algorithme de calcul pour données réelles.

Une autre méthode pour calculer les transformées de nombres réels consiste à faire appel aux transformées impaires [5].

La Transformée de Fourier Discrète impaire établit par définition les relations suivantes entre deux ensembles de N nombres complexes $x(n)$ et $X(k)$:

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi \frac{(2k+1)n}{2N}} \tag{3.18}$$

$$x(n) = \sum_{k=0}^{N-1} X(k)\, e^{j2\pi \frac{(2k+1)n}{2N}}$$

Les coefficients de cette transformée ont pour affixe les points M du cercle unité tels que le vecteur $\overrightarrow{OM}$ fasse avec l'axe des abscisses un angle multiple impair de $\frac{2\pi}{2N}$ comme le montre la figure 3.1.

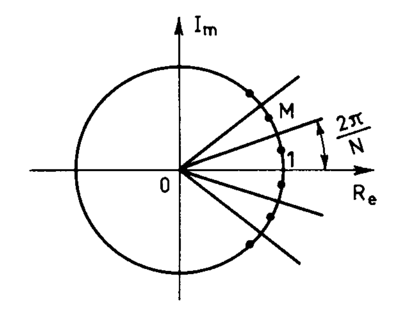


FIG. 3.1. *Coefficients de la TFD impaire*

En posant $W = e^{-j\frac{\pi}{N}}$ la matrice de cette transformée s'écrit :

$$T_N^1 = \begin{bmatrix} 1 & W & W^2 & \dots & W^{N-1} \\ 1 & W^3 & W^6 & \dots & W^{3(N-1)} \\ 1 & W^5 & W^{10} & \dots & W^{5(N-1)} \\ \vdots & \vdots & & \dots & \vdots \\ 1 & W^{(2N-1)} & \dots\dots\dots & \dots & W^{(2N-1)(N-1)} \end{bmatrix}$$

Si les $x(n)$ sont des nombres réels, on peut écrire :

$$\begin{aligned} X(N-1-k) &= \frac{1}{N}\sum_{n=0}^{N-1} x(n)\, e^{-j2\pi\frac{2(N-1-k)+1}{2N}} \\ &= \frac{1}{N}\sum_{n=0}^{N-1} x(n)\, e^{j2\pi\frac{2k+1}{2N}n} \end{aligned} \tag{3.19}$$

D'où le résultat suivant :

$$X(N-1-k) = \overline{X(k)} \tag{3.20}$$

Par suite les $X(k)$ d'indice pair et d'indice impair sont complexes conjugués ; il suffit donc de calculer les $X(k)$ d'indice pair pour faire une transformée sur des réels. Une telle transformée correspond à la matrice T_R telle que :

$$T_R = \begin{bmatrix} 1 & W & W^2 & \dots & W^{\frac{N}{2}} & \dots & W^{N-1} \\ 1 & W^5 & W^{10} & \dots & W^{\frac{5N}{2}} & \dots & W^{5(N-1)} \\ \vdots & \vdots & & & \vdots & & \vdots \\ 1 & W^{2N-3} & \dots\dots\dots & & W^{\frac{(2N-3)N}{2}} & \dots & W^{(2N-3)(N-1)} \end{bmatrix}$$

Soit $D_{\frac{N}{2}}$ la matrice diagonale dont les éléments de la diagonale sont les W^k avec $0 \leqslant k \leqslant \frac{N}{2} - 1$ et $T_{\frac{N}{2}}$ la matrice de la transformée d'ordre $\frac{N}{2}$. Compte tenu du fait que $W^{2N} = 1$, et $W^{\frac{N}{2}} = -j$, il vient :

$$T_R = [T_{\frac{N}{2}} D, -jT_{\frac{N}{2}} D] = (T_{\frac{N}{2}} D) \otimes [1 \; -j] \tag{3.21}$$

Alors la transformée impaire sur des nombres réels se calcule en effectuant une transformée d'ordre $\frac{N}{2}$ sur la suite des nombres complexes

$$y(n) = \left[x(n) - jx\left(\frac{N}{2} + n\right)\right] W^n \quad \text{avec} \quad 0 \leqslant n \leqslant \frac{N}{2} - 1. \tag{3.22}$$

Le nombre de calculs est le même que dans la méthode indiquée au début du paragraphe, mais la structure est plus simple. Il faut noter cependant que les nombres transformés donnent un échantillonnage du spectre du signal que représente les $x(n)$, décalé d'un demi-pas d'échantillonnage en fréquence.

Un cas important où des simplifications notables interviennent est celui des suites réelles symétriques. Les réductions de calcul sont mises en évidence par utilisation de la transformée doublement impaire [6].

3.3.2 La Transformée doublement impaire

La Transformée de Fourier Discrète doublement Impaire établit par définition les relations suivantes entre deux ensembles de N nombres complexes $x(n)$ et $X(k)$:

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-j2\pi \frac{(2k+1)(2n+1)}{4N}} \tag{3.23}$$

$$x(n) = \sum_{k=0}^{N-1} X(k) e^{j2\pi \frac{(2k+1)(2n+1)}{4N}} \tag{3.24}$$

Les coefficients de cette transformée ont pour affixe les points M du cercle unité tels que le vecteur $\overrightarrow{OM}$ fasse avec l'axe des abscisses un angle multiple impair de $\frac{2\pi}{4N}$ comme le montre la figure 3.2.

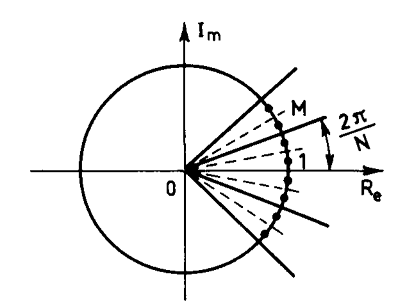


FIG. 3.2. *Coefficients de la TFD doublement impaire*

Si les $x(n)$ sont des nombres réels on vérifie la relation :

$$X(N-1-k) = -\overline{X}(k)$$

De même si les $X(k)$ sont réels alors :

$$x(N-1-n) = -\bar{x}(n)$$

En posant comme précédemment $W = e^{-j\frac{\pi}{N}}$ la matrice de la transformée s'écrit :

$$T_N^{II} = \begin{bmatrix} W^{\frac{1}{2}} & & W^{\frac{5}{2}} & \dots & W^{N-\frac{1}{2}} \\ W^{\frac{3}{2}} & W^{\frac{9}{2}} & W^{\frac{15}{2}} & \dots & W^{3\left(N-\frac{1}{2}\right)} \\ W^{\frac{5}{2}} & W^{\frac{15}{2}} & W^{\frac{25}{2}} & \dots & W^{5\left(N-\frac{1}{2}\right)} \\ \vdots & \vdots & \vdots & & \vdots \\ W^{N-\frac{1}{2}} & W^{3\left(N-\frac{1}{2}\right)} & & \dots\dots & W^{(2N-1)\left(N-\frac{1}{2}\right)} \end{bmatrix} \quad (3.25)$$

Une telle transformée se factorise comme suit :

$$T_N^{II} = W^{\frac{1}{2}} \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & W & 0 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & & & W^{N-1} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & W^2 & W^4 & \dots & W^{2(N-1)} \\ 1 & W^4 & W^8 & \dots & W^{4(N-1)} \\ \vdots & \vdots & & & \vdots \\ 1 & W^{2(N-1)} & & \dots & W^{2(N-1)(N-1)} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & W & 0 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & & \dots & W^{N-1} \end{bmatrix}$$

C'est-à-dire que :

$$T_N^{II} = W^{\frac{1}{2}} D_N T_N D_N \quad (3.26)$$

Considérons le cas où la suite des données $x(n)$ est une suite réelle et antisymétrique, c'est-à-dire que $x(n) = -x(N-1-n)$; alors il en est de même de la suite $X(k)$. La suite des $x(n)$ pour n pair est égale à la suite des $x(n)$ pour n impair au signe près ; de même pour la suite des $X(k)$.

Pour calculer la transformée il suffit dans ce cas de mener les calculs sur les $x(2n)$ avec $0 \leqslant n \leqslant \frac{N}{2} - 1$ puisque les $X(k)$ sont des nombres réels ; d'autre part il suffit de faire ces calculs sur les $X(2k)$ avec $0 \leqslant k \leqslant \frac{N}{2} - 1$.

La matrice correspondante T_{RR} s'écrit :

$$T_{RR} = W^{\frac{1}{2}} \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & W^2 & 0 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & & \dots\dots & W^{2\left(\frac{N}{2}-1\right)} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & W^2 & \cdots & W^{8\left(\frac{N}{2}-1\right)} \\ 1 & W^{16} & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & W^{8\left(\frac{N}{2}-1\right)} & & W^{8\left(\frac{N}{2}-1\right)\left(\frac{N}{2}-1\right)} \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & W^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & W^{2\left(\frac{N}{2}-1\right)} \end{bmatrix}$$

Compte tenu du fait que $W^{2N} = 1$, il vient :

$$T_{RR} = W^{\frac{1}{2}} D_{\frac{N}{2}} \begin{bmatrix} T_{\frac{N}{4}} & T_{\frac{N}{4}} \\ T_{\frac{N}{4}} & T_{\frac{N}{4}} \end{bmatrix} D_{\frac{N}{2}} \qquad (3.27)$$

Et comme $W^{2,\frac{N}{4}} = -j$ ce calcul peut être conduit en n'effectuant qu'une seule fois les opérations représentées par la matrice $T_{\frac{N}{4}}$ sur l'ensemble de nombres $x(2n) - jx\left(2n + \frac{N}{2}\right)$ avec $0 \leqslant n \leqslant \frac{N}{4} - 1$. Les $\frac{N}{4}$ nombres obtenus sont des nombres complexes dont les parties réelles constituent l'ensemble des $X(2k)$ avec $0 \leqslant k \leqslant \frac{N}{4} - 1$ recherchées. En effectuant l'opération définie par T_{RR} pour les nombres transformés de rang $2k + \frac{N}{2}$ avec $0 \leqslant k \leqslant \frac{N}{4} - 1$, on peut vérifier que l'on obtient les nombres précédents multipliés par $-j$; c'est-à-dire que la partie imaginaire des nombres obtenus précédemment, fournit l'ensemble des $X\left(2k + \frac{N}{2}\right)$. Il s'en suit que la transformée doublement impaire appliquée à une suite de N termes, réelle antisymétrique, ou à une suite symétrique rendue antisymétrique par un changement de signe convenable, se réduit à l'équation :

$$\left[X(2k) + jX\left(2k + \frac{N}{2}\right)\right] = W^{\frac{1}{2}} D_{\frac{N}{4}} T_{\frac{N}{4}} D_{\frac{N}{4}} \left[x(2n) - jx\left(2n + \frac{N}{2}\right)\right] \qquad (3.28)$$

avec $0 \leqslant k \leqslant \frac{N}{4} - 1$, $0 \leqslant n \leqslant \frac{N}{4} - 1$ et où $D_{\frac{N}{4}}$ est une matrice diagonale dont les éléments sont les W^{2i} avec $0 \leqslant i \leqslant \frac{N}{4} - 1$.

Le nombre de multiplications complexes nécessaire Mc s'élève à :

$$Mc = \frac{N}{8} \log 2\left(\frac{N}{8}\right) + 2\,\frac{N}{4} = \frac{N}{8} \log 2\,(2N) \qquad (3.29)$$

Les comparaisons de volume de calculs entre les différentes transformées conduisent au tableau 3.1 :

Tableau 3.1 – Quantités de calculs dans diverses TFR.

	Multiplications complexes	Additions complexes	Positions de mémoire
TFD complexe	$\frac{N}{2}\log_2\left(\frac{N}{2}\right)$	$N\log_2(N)$	2N
TFD impaire-données réelles	$\frac{N}{4}\log_2(N)$	$\frac{N}{2}\log_2\left(\frac{N}{2}\right)$	N
TFD doublement impaire-données réelles paires	$\frac{N}{8}\log_2(2N)$	$\frac{N}{4}\log_2\left(\frac{N}{4}\right)$	$\frac{N}{2}$

L'intérêt des transformées impaires apparaît clairement. Il faut cependant noter que d'autres algorithmes permettent d'obtenir avec les données réelles et réelles symétriques des réductions en calcul un peu supérieures [7], mais sans avoir la facilité de mise en œuvre, en particulier pour les réalisations matérielles, qu'offrent les transformées impaires.

Une particularité de la transformée doublement impaire appliquée à une suite réelle antisymétrique est qu'elle est identique à la transformée inverse ; il n'y a pas de distinction, mis à part le facteur d'échelle $\frac{1}{N}$, entre transformées directe et inverse dans ce cas.

La transformée de Fourier d'une suite réelle symétrique intervient par exemple dans le calcul de la densité spectrale énergétique d'un signal à partir de la fonction d'autocorrélation.

3.3.3 Les Transformées discrètes en cosinus et sinus

Les transformées considérées jusqu'à présent ont des coefficients complexes. Des transformées discrètes de la même famille peuvent être obtenues à partir des parties réelles et imaginaires des coefficients complexes. On peut ainsi définir :

- Une transformée de Fourier Discrète en cosinus (TFD-cos) :

$$X_{FC}(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\cos\left(\frac{2\pi nk}{N}\right) \tag{3.30}$$

- Une transformée de Fourier Discrète en sinus (TFD-sin) :

$$X_{FS}(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\sin\left(\frac{2\pi nk}{N}\right) \tag{3.31}$$

- Une transformée en cosinus discrète (TCD) :

$$X_{CD}(0) = \frac{\sqrt{2}}{N} \sum_{n=0}^{N-1} x(n)$$

$$X_{CD}(k) = \frac{2}{N} \sum_{n=0}^{N-1} x(n) \cos\left(\frac{2\pi(2n+1)k}{4N}\right) \qquad (3.32)$$

à laquelle correspond la transformée inverse :

$$x(n) = \frac{1}{\sqrt{2}} X_{CD}(0) + \sum_{k=1}^{N-1} X_{CD}(k) \cos\left(\frac{2\pi(2n+1)k}{4N}\right)$$

- Une transformée en sinus discrète (TSD) :

$$X_{SD}(k) = \sqrt{\frac{2}{N+1}} \sum_{n=0}^{N-1} x(n) \sin\left(\frac{2\pi(n+1)(k+1)}{2N+2}\right) \qquad (3.33)$$

À l'aide de manipulations comparables à celles des données dans les paragraphes précédents, on peut établir des relations entre la TFD et ces diverses transformées ainsi qu'entre ces transformées elles-mêmes.

Par exemple, d'après les définitions, il vient :

$$\text{TFD}(N) = \text{TFD-cos}(N) - j\,\text{TFD-sin}(N)$$

Considérant la transformée en cosinus, il vient :

$$X_{FC}(k) = \sum_{n=0}^{N/2-1} x(2n) \cos\left(\frac{2\pi nk}{N/2}\right) + \sum_{n=0}^{N/4-1} [x(2n+1) + x(N-2n-1)] \cos\left(\frac{2\pi(2n+1)k}{4.N/4}\right)$$

c'est-à-dire que la transformée en cosinus d'ordre N peut se calculer à l'aide d'une transformée en cosinus d'ordre $\frac{N}{2}$ et d'une transformée discrète d'ordre N/4, soit sous forme concise :

$$\text{TFD-cos}(N) = \text{TFD-cos}\left(\frac{N}{2}\right) + \text{TCD}\left(\frac{N}{4}\right)$$

De même la transformée en cosinus discrète (TCD) s'écrit :

$$X_{CD}(k) = \frac{2}{N} \sum_{n=0}^{N/2-1} \left[x(2n) \cos \frac{2\pi(4n+1)k}{4N} + x(2n+1) \cos \frac{2\pi[4(N-n-1)+1]k}{4N}\right.$$

c'est-à-dire qu'en posant, pour $0 \leqslant n \leqslant N/2-1$:

$$y(n) = x(2n)$$
$$y(N-n-1) = x(2n+1)$$

il vient :

$$X_{CD}(k) = \frac{2}{N} \sum_{n=0}^{N-1} y(n) \cos \frac{2\pi(4n+1)k}{4N}$$

et en développant le cosinus :

$$\text{TCD}(N) = \cos \frac{2\pi k}{4N} \text{TFD-cos}(N) - \sin \frac{2\pi k}{4N} \text{TFD-sin}(N)$$

ce qui s'écrit également, en fonction des données et sous une forme concise :

$$\text{TCD}(x) = 2c(k) \, \text{R}e \left\{ e^{-j\frac{\pi k}{2N}} \text{TFD}(y) \right\} \tag{3.34}$$

avec $c(0) = \frac{1}{\sqrt{2}}$ et $c(k) = 1$ pour $k = 1, \ldots, N-1$.

Ainsi, la transformée en cosinus discrète d'ordre N peut se calculer à l'aide d'une transformée de Fourier discrète de même ordre. Compte tenu du fait que seule la partie réelle est utilisée dans l'expression ci-dessus, on peut même calculer 2 TCD, en utilisant également la partie imaginaire. La même méthode s'applique à la transformée inverse et on peut calculer 2 TCD inverses avec une TFD inverse, en effectuant les opérations indiquées à la figure 3.3. Les relations entre les variables :

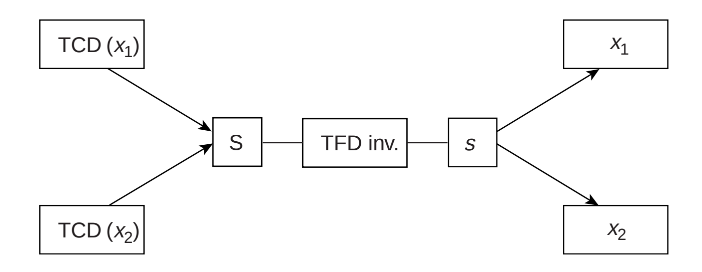


FIG. 3.3 *Calcul de 2 TCD inverses avec une TFD inverse de même dimension*

à l'entrée de la TFD inverse sont les suivantes [8] :

$$S_0 = \frac{C_0(x_1) + jC_0(x_2)}{\sqrt{2}} \; ; \; S_{N/2} = \frac{C_{N/2}(x_1) + jC_{N/2}(x_2)}{\sqrt{2}}$$

$$2\,S_k = \left\{ [C_k(x_1) + C_{N-k}(x_2)] \cos\left(\frac{\pi k}{2N}\right) + [C_{N-k}(x_1) - C_{N-k}(x_2)] \sin\left(\frac{\pi k}{2N}\right) \right\}$$
$$+ j \left\{ [C_k(x_1) + C_{N-k}(x_2)] \sin\left(\frac{\pi k}{2N}\right) + [C_k(x_2) - C_{N-k}(x_1)] \cos\left(\frac{\pi k}{2N}\right) \right.$$

avec $k = 1, \ldots, N-1$; $k \neq N/2$.

De même en sortie de la TFD inverse :

$$x_1(2p) = \text{R}e\,\{s(p)\} ; x_1(2p+1) = \text{R}e\,\{s(N-p-1)\} ;$$
$$x_2(2p) = \text{I}m\,\{s(p)\} ; x_2(2p+1) = \text{I}m\,\{s(N-p-1)\}$$

La méthode permet de réduire la quantité de calculs en compression d'images, par exemple.

Parmi les transformées à coefficients réels, on peut aussi mentionner la transformée, dite de Hartley discrète (THD), définie par :

$$X_{HD}(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) \left[\cos 2\pi \frac{nk}{N} + \sin 2\pi \frac{nk}{N} \right] \tag{3.35}$$

et pour la transformée inverse :

$$x(n) = \sum_{k=0}^{N-1} X_{HD}(k) \left[\cos 2\pi \frac{nk}{N} + \sin 2\pi \frac{nk}{N} \right]$$

La liaison avec la TFD est donnée par [9] :

$$X(k) = \frac{1}{2} [X_{HD}(k) + X_{HD}(N-1-k) - j(X_{HD}(k) - X_{HD}(N-1-k))] \tag{3.36}$$

La transformée en cosinus discrète est utilisée en compression de l'informatique, notamment en traitement d'images. En effet, elle fournit pour les signaux d'images une approximation de la transformation propre, qui permet de représenter un signal par le minimum de composantes.

Ce pouvoir de compression provient du fait qu'elle élimine les discontinuités de bord de la TFD mentionnées au paragraphe 2.1, car elle correspond à la TFD d'une suite symétrisée. Pour pouvoir effectuer cette symétrisation avant d'appliquer la TFD, il faut éviter d'avoir une valeur à l'indice zéro, ce qui est obtenu en prenant la suite $u(n)$ telle que :

$$u(2p) = 0 \quad ; \quad 0 \leqslant p \leqslant 2N-1$$
$$u(2p+1) = u(4N-2p-1) = x(p) \quad ; \quad 0 \leqslant p \leqslant N-1$$

La TFD d'ordre 4N de la suite $u(n)$ conduit, après simplifications, à l'expression (3.32).

3.3.4 La transformée en cosinus discrète à 2 dimensions

La transformée en cosinus discrète à deux dimensions (TCD-2D) est définie, pour un ensemble de $N \times N$ données réelles, par les équations :

$$X(k_1, k_2) = \frac{4e(k_1)e(k_2)}{N^2} \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} x(n_1, n_2) \cos \frac{2\pi(2n_1+1)k_1}{4N} \cos \frac{2\pi(2n_2+1)k_2}{4N} \tag{3.37}$$

et :

$$x(n_1, n_2) = \sum_{k_1=0}^{N-1} \sum_{k_2=0}^{N-1} e(k_1)\, e(k_2)\, X(k_1, k_2) \cos\left(\frac{2\pi(2n_1+1)k_1}{4N}\right) \cos\left(\frac{2\pi(2n_2+1)k_2}{4N}\right)$$

avec :

$$e(k) = \frac{1}{\sqrt{2}} \quad \text{pour} \quad k = 0$$

$$e(k) = 1 \quad \text{pour} \quad k \neq 0.$$

Cette transformée est séparable et peut être calculée de la manière suivante :

$$\mathrm{X}(k_1, k_2) = \frac{2}{\mathrm{N}} e(k_2) \sum_{n_2=0}^{\mathrm{N}-1} \cos \frac{2\pi(2n_2+1)k_2}{4\mathrm{N}} \left[\frac{2}{\mathrm{N}} e(k_1) \sum_{n_1=0}^{\mathrm{N}-1} x(n_1, n_2) \cos \frac{2\pi(2n_1+1)k_1}{4\mathrm{N}} \right]$$

Ainsi la transformée à deux dimensions se calcule en utilisant 2N fois l'algorithme correspondant à la transformée TCD à une dimension et le nombre de multiplications réelles à effectuer est de l'ordre de $\mathrm{N}^2 \log 2\,(\mathrm{N})$. En fait, cette valeur peut être atteinte, notamment en faisant appel à un algorithme basé sur la réduction d'une TCD d'ordre N à deux TCD d'ordre $\frac{\mathrm{N}}{2}$ [10]. Il est même possible d'atteindre la valeur $\frac{3}{4}\mathrm{N}^2 \log 2\,(\mathrm{N})$ en n'utilisant pas la propriété de séparatibilité de la TCD-2D mais en étendant à deux dimensions le principe d'entrelacement par décomposition de l'ensemble des $\mathrm{N} \times \mathrm{N}$ données en ensembles de $\frac{\mathrm{N}}{2} \times \frac{\mathrm{N}}{2}$ données [11].

Appliquée à une image, la TCD-2D fait apparaître les fréquences spatiales. Par exemple, pour une ligne verticale définie par :

$$x(n_1, n_2) = \frac{1}{\mathrm{N}}; \quad n_1 = 0$$

$$x(n_1, n_2) = 0; \quad n_1 \neq 0$$

les valeurs transformées $\mathrm{X}(k_1, k_2)$ s'annulent pour $k_2 \neq 0$. On vérifie également qu'une diagonale constante est transformée en une diagonale constante, la matrice unité est un élément propre de la transformation.

3.4 TRANSFORMÉE AVEC RECOUVREMENT

La fonction de filtrage de la TFD peut être améliorée en considérant des transformées effectuées sur des blocs de données qui se recouvrent [12].

Soit un bloc de données de longueur 2M, double de l'ordre M de la transformée. Au temps n, le bloc de données traitées est représenté par un ensemble de 2 vecteurs à M éléments désignés par $\mathrm{X}_1(n)$ et $\mathrm{X}_2(n)$ comme le montre la figure 3.4.

Fig. 3.4. *Recouvrement des blocs de données*

Au temps $(n + 1)$, le bloc de données comporte la moitié des données du bloc précédent, c'est-à-dire que $X_1(n + 1) = X_2(n)$.

La transformée avec recouvrement permet de restituer les vecteurs $X_1(n)$ et $X_2(n)$. Sa matrice comporte 2M lignes et M colonnes et elle correspond aux opérations suivantes, sur deux blocs consécutifs :

$$U_1 = [A, B]\begin{bmatrix} X_1(n) \\ X_2(n) \end{bmatrix}$$

et (3.38)

$$U_2 = [A, B]\begin{bmatrix} X_1(n+1) \\ X_2(n+1) \end{bmatrix}$$

où A et B sont deux matrices carrées de côté M.

Si l'on considère maintenant les opérations :

$$\begin{bmatrix} Y_1(n) \\ Y_2(n) \end{bmatrix} = \begin{bmatrix} A^t \\ B^t \end{bmatrix} U_1 ; \quad \begin{bmatrix} Y_1(n+1) \\ Y_2(n+1) \end{bmatrix} = \begin{bmatrix} A^t \\ B^t \end{bmatrix} U_2 \tag{3.39}$$

on obtient :

$$\frac{1}{2}[Y_2(n) + Y_1(n+1)]$$

$$= [B^tAX_1(n) + B^tBX_2(n) + A^tAX_1(n+1) + A^tBX_2(n+1)]\frac{1}{2} \tag{3.40}$$

Pour retrouver les données d'origine à la fin du temps $(n + 1)$, c'est-à-dire $X_2(n) = \frac{1}{2}[Y_2(n) + Y_1(n+1)]$, il faut que les conditions suivantes soient vérifiées :

$$B^tA = A^tB = 0$$

$$\frac{1}{2}[B^tB + A^tA] = I_M \tag{3.41}$$

Par exemple, ces conditions sont vérifiées si les éléments de la matrice [A, B] de la transformée sont tels que :

$$a_{nk} = h(n)\sqrt{\frac{2}{M}}\cos\left[\left(n + \frac{M+1}{2}\right)\left(k + \frac{1}{2}\right)\frac{\pi}{M}\right]$$

avec $0 \leqslant n \leqslant 2M - 1 ; \quad 0 \leqslant k \leqslant M - 1$ et

$$h(n) = -\sin\left(n + \frac{1}{2}\right)\frac{\pi}{2M}$$

En fait, on a obtenu un banc de M filtres orthogonaux et les termes $h(n)$ sont les coefficients du filtre prototype qui, dans l'exemple est un filtre passe-bas en sinus dont la réponse en fréquence, donnée au paragraphe 5.8, est plus sélective que celle du filtre de la TFD.

En compression d'image, les transformées avec recouvrement produisent un lissage et permettent de réduire les effets dits de blocage.

3.5 AUTRES ALGORITHMES DE CALCUL RAPIDE

Les algorithmes de TFR constituent une technique pour calculer une TFD d'ordre N avec un nombre de multiplications de l'ordre de N log 2(N). Dans les paragraphes précédents, il a été montré que ces algorithmes ont une structure relativement simple et offrent suffisamment de souplesse pour qu'une bonne adaptation aux contraintes d'exploitation et aux caractéristiques technologiques puisse être aisément atteinte, d'où leur grand intérêt pratique.

Cependant ils ne constituent pas la seule méthode de calcul rapide de la TFD et l'on peut élaborer des algorithmes qui nécessitent, tout au moins dans certains cas, un temps de calcul moindre ou un nombre de multiplications plus faible, ou qui sont applicables quel que soit l'ordre N et pas obligatoirement une puissance de deux.

Une première approche consiste à remplacer les multiplications complexes, coûteuses en circuits ou en temps de machine, par un ensemble d'opérations plus faciles à mettre en œuvre. La référence [13] décrit une technique qui utilise une caractéristique de la TFD mentionnée au paragraphe 2.4.1, le fait que les opérations de multiplication par les coefficients W^k, correspondent à des rotations de phase. La technique dite CORDIC (calcul numérique à rotation des coordonnées) permet de réaliser ces rotations par un enchaînement d'opérations simples : pour faire tourner un vecteur (x, y) d'un angle θ avec la précision $\frac{\theta}{2^n}$, on opère une suite de n rotations élémentaires d'angles $d\theta_i$ tels que $\text{tg}\, d\theta_i = 2^{-i}$ avec $0 \leqslant i \leqslant n - 1$ et $-\frac{\pi}{2} \leqslant \theta \leqslant \frac{\pi}{2}$. Les coordonnées x_i et y_i du vecteur à l'itération i conduisent aux coordonnées à l'itération $i + 1$ par les relations :

$$x_{i+1} = x_i + \text{signe}\,[\theta_i]\,.\,y_i 2^{-i}$$
$$y_{i+1} = y_i - \text{signe}\,[\theta_i]\,.\,x_i 2^{-i} \qquad (3.42)$$
$$\theta_{i+1} = \theta_i - \text{signe}\,[\theta_i]\,.\,d\theta_i$$

La fonction signe $[\theta_i]$ est le signe de θ_i et on prend $\theta_0 = -\theta$. Ces opérations ne comportent que des additions avec décalages, elles peuvent être plus avantageuses que la multiplication complexe de même précision.

Le calcul d'une TFD d'ordre N avec un volume de multiplications qui soit de l'ordre de N, au lieu de N log 2(N), peut être obtenu par une factorisation de la matrice T_N d'un type particulier. En effet la matrice T_N peut se décomposer en un produit de trois facteurs :

$$T_N = B_N C_N A_N$$

où A_N est une matrice de dimension $J \times N$, avec J entier, C_N est une matrice diagonale de dimension J et B_N une matrice de dimension $N \times J$. La particularité de cette factorisation réside dans le fait que les éléments des matrices A_N et B_N sont 0, 1 ou – 1. Dans ces conditions le calcul demande J multiplications. Par exemple la multiplication complexe se met sous cette forme :

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} a & 0 & 0 \\ 0 & a+b & 0 \\ 0 & 0 & a-b \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

ce qui permet de l'effectuer avec 3 multiplications réelles seulement comme indiqué au paragraphe 2.7. Cette décomposition est évidente pour $J = N^2$, par exemple pour N = 3, il vient :

$$T_3 = \begin{bmatrix} 1&1&1&0&0&0&0&0&0 \\ 0&0&0&1&1&1&0&0&0 \\ 0&0&0&0&0&0&1&1&1 \end{bmatrix} \begin{bmatrix} 1 &&&&&&&& \\ & 1 &&&&& \text{O} && \\ && 1 &&&&&& \\ &&& 1 &&&&& \\ &&&& W &&&& \\ &&&&& W^2 &&& \\ &&&&&& 1 && \\ & \text{O} &&&&&& W^2 & \\ &&&&&&&& W \end{bmatrix} \begin{bmatrix} 1&0&0 \\ 0&1&0 \\ 0&0&1 \\ 1&0&0 \\ 0&1&0 \\ 0&0&1 \\ 1&0&0 \\ 0&1&0 \\ 0&0&1 \end{bmatrix}$$

Avec certaines valeurs de N faibles, il existe des factorisations telles que J soit de l'ordre de N, alors il en est de même du nombre de multiplications. Pour généraliser cette propriété et mettre en évidence une factorisation de T_N convenable il faut opérer une permutation des données avant et après transformation. Par exemple, pour N = 12, en posant :

$$X' = \begin{bmatrix} X_0 \\ X_3 \\ X_6 \\ X_9 \\ X_4 \\ X_7 \\ X_{10} \\ X_1 \\ X_8 \\ X_{11} \\ X_2 \\ X_5 \end{bmatrix} \quad \text{et} \quad x' = \begin{bmatrix} X_0 \\ X_9 \\ X_6 \\ X_3 \\ X_4 \\ X_1 \\ X_{10} \\ X_7 \\ X_8 \\ X_5 \\ X_2 \\ X_{11} \end{bmatrix}$$

et en faisant appel aux produits de Kronecker de matrices, on vérifie que :

$$X' = (T_3 \otimes T_4)x'$$

De même, si N se décompose en L facteurs premiers entre eux :

$$N = N_L . N_{L-1} \dots N_1$$

On peut montrer que :

$$X' = (T_{N_L} \otimes T_{N_{L-1}} \otimes \dots \otimes T_{N_1})x' \tag{3.43}$$

En utilisant la factorisation définie précédemment pour les matrices T_{N_L} et les propriétés algébriques des produits de Kronecker, il vient :

$$X' = (B_{N_L} \otimes B_{N_{L-1}} \otimes \dots \otimes B_{N_1})(C_{N_L} \otimes C_{N_{L-1}} \otimes \dots C_{N_1})$$
$$(A_{N_L} \otimes A_{N_{L-1}} \otimes \dots \otimes A_{N_1})x'$$

Ce résultat définit un type d'algorithme appelé algorithme de Winograd.

Il apparaît clairement que l'algorithme d'ordre N se déduit des algorithmes d'ordre N_i avec $1 \leqslant i \leqslant L$; d'où l'importance des algorithmes à faible nombre de multiplications pour les petites valeurs de N. La référence [14] donne des algorithmes pour N = 2, 3, 4, 5, 7, 8, 9, 16, où le nombre de multiplications est de l'ordre de N, comme le montre le tableau 3.2. Dans la colonne des multiplications, les chiffres entre parenthèses donnent le nombre de multiplications par des coefficients différents de 1. De plus ces multiplications sont des multiplications complexes qui correspondent à deux multiplications réelles. Le nombre d'additions est comparable à celui des algorithmes de TFR.

Les algorithmes pour les faibles valeurs de N sont obtenus en calculant la Transformée de Fourier comme un ensemble de corrélations :

$$X_k = \sum_{n=1}^{N-1} (x_n - x_0) W^{nk}; \quad k = 1, \dots, N-1$$

Tableau 3.2. – NOMBRE D'OPÉRATIONS DANS LES ALGORITHMES DE WINOGRAD D'ORDRE N FAIBLE.

Ordre de la TFD	Multiplications	Additions
2	2 (0)	2
3	3 (2)	6
4	4 (0)	8
5	6 (5)	17
7	9 (8)	36
8	8 (2)	26
9	11 (10)	44
16	18 (10)	74

et en utilisant les propriétés algébriques de l'ensemble des exposants de W, définis modulo N.

Par exemple, pour N = 4, l'enchaînement des opérations est le suivant :

$$t_1 = x_0 + x_2, \qquad t_2 = x_1 + x_3$$
$$m_0 = 1\,.(t_1 + t_2), \quad m_1 = 1\,.(t_1 - t_2)$$
$$m_2 = 1\,.(x_0 - x_2), \quad m_3 = j\,(x_1 - x_3)$$
$$X_0 = m_0$$
$$X_1 = m_2 + m_3$$
$$X_2 = m_1$$
$$X_3 = m_2 - m_3$$

Pour N = 8 :

$$t_1 = x_0 + x_4, \quad t_2 = x_2 + x_6, \quad t_3 = x_1 + x_5,$$
$$t_4 = x_1 - x_5, \quad t_5 = x_3 + x_7, \quad t_6 = x_3 - x_7,$$
$$t_7 = t_1 + t_2, \quad t_8 = t_3 + t_5,$$

$$m_0 = 1\,.(t_7 + t_8), \qquad m_1 = 1\,.(t_7 - t_8),$$
$$m_2 = 1\,.(t_1 - t_2), \qquad m_3 = 1\,.(x_0 - x_4),$$
$$m_4 = \cos\left(\frac{\pi}{4}\right).(t_4 - t_6), \quad m_5 = j\,(t_3 - t_5),$$
$$m_6 = j\,.(x_2 - x_6), \qquad m_7 = j\sin\left(\frac{\pi}{4}\right).(t_4 + t_6),$$

$$s_1 = m_3 + m_4, \quad s_2 = m_3 - m_4, \quad s_3 = m_6 + m_7 \quad s_4 = m_6 - m_7$$
$$X_0 = m_0, \quad X_1 = s_1 + s_3, \quad X_2 = m_2 + m_5, \quad X_3 = s_2 - s_4$$
$$X_4 = m_1, \quad X_5 = s_2 + s_4, \quad X_6 = m_2 - m_5, \quad X_7 = s_1 - s_3.$$

Finalement les algorithmes de Winograd apportent une réduction du volume de calcul, qui peut être importante, par rapport aux algorithmes de TFR. Il en est de même pour d'autres algorithmes, comme ceux qui consistent à utiliser les transformées polynomiales [15].

Ces techniques peuvent apparaître intéressantes dans certaines applications, mais il faut bien noter qu'elles peuvent conduire à une plus grande capacité de mémoire et à un enchaînement plus compliqué des opérations, se traduisant par un accroissement du volume de matériel de l'unité de commande du système ou de la taille des mémoires de programme.

Une autre voie séduisante dans la recherche de l'optimisation du traitement et des machines, est celle qui fait appel aux transformations algébriques.

3.6 TRANSFORMÉE DE FOURIER BINAIRE – HADAMARD

La transformée de Fourier discrète et ses variantes nécessitent une quantité de calculs qui peut être jugée excessive pour certaines applications, certains traitements d'images en temps réel par exemple. Alors, on peut faire appel à des transformées ayant des propriétés comparables, mais sans multiplications, comme la transformée dite de Fourier binaire ou Hadamard [16].
La transformée de Hadamard d'ordre $N = 2^M$ est définie par la matrice H_N qui se déduit de la TFD d'ordre 2, T_2, par produits de Kronecker :

$$H_2 = T_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} ; H_4 = T_2 \otimes T_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} ;$$

$$H_N = T_2 \otimes H_{N/2} \tag{3.44}$$

C'est une transformée réelle, symétrique et orthogonale :

$$\frac{1}{N} H_N H_N = I_N$$

Les algorithmes rapides sont les mêmes que pour la TFR, avec les croisillons mais sans les multiplications. Sur le plan du filtrage, le banc de filtres obtenu est beaucoup moins sélectif que celui de la TFD, car les fonctions élémentaires comportent une fréquence et des harmoniques. Par exemple, la figure 3.5 représente le module de la transformée de Fourier de la ligne 31 de la matrice H_{64}.

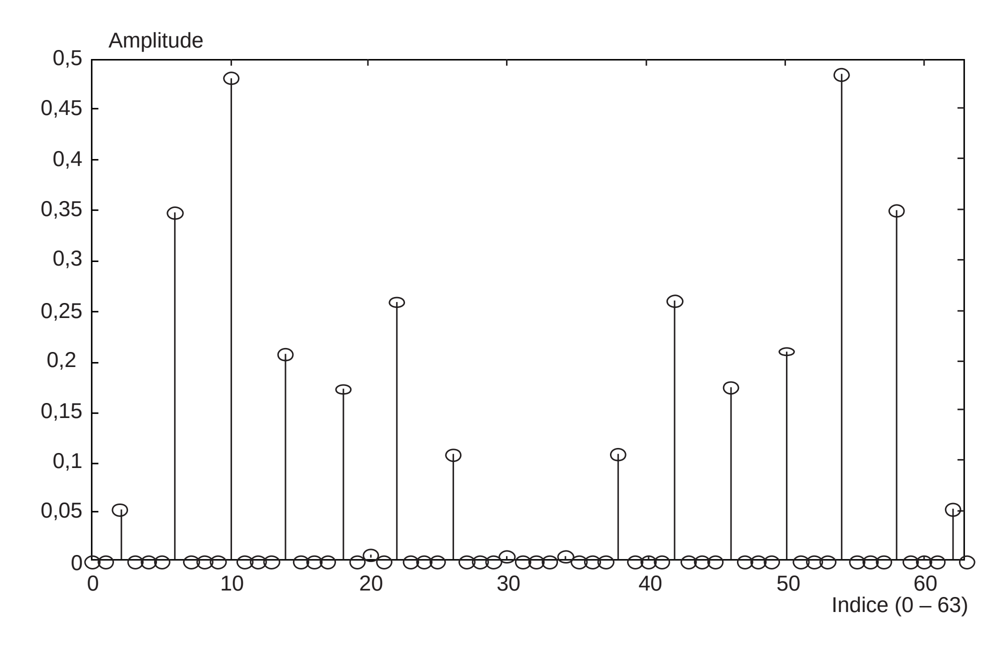


Fig. 3.5. *Spectre du code 31/64 de Hadamard*

Les matrices de Hadamard trouvent un domaine d'application important en radiocommunications mobiles, car les éléments sont utilisés comme codes d'étalement dans les systèmes basés sur l'étalement de spectre.

3.7 LES TRANSFORMATIONS ALGÉBRIQUES

La transformation de Fourier conduit à faire des opérations arithmétiques dans le corps des nombres complexes ; les machines qui réalisent ces opérations utilisent généralement des représentations binaires qui sont des approximations des données et des coefficients. La précision des calculs est fonction du nombre de bits disponible dans la machine.

En fait une machine à B bits effectue des opérations sur l'ensemble à 2^B éléments des entiers : 0, 1, …, $2^B - 1$. Dans cet ensemble les lois d'addition et multiplication habituelles ne peuvent être respectées, il faut introduire le décalage et la troncation qui entraînent des approximations dans les calculs, comme indiqué au paragraphe 2.3.

La première condition à remplir pour que les calculs soient exacts dans un ensemble E est que le produit ou la somme de deux éléments de l'ensemble E appartiennent à cet ensemble. Cette condition est vérifiée dans l'ensemble des entiers 0, 1, …, $M - 1$ si les calculs sont faits modulo M. En choisissant convenablement le module M il est possible de définir des transformations ayant des propriétés comparables à celles de la TFD et permettant le calcul de convolutions sans erreur avec des algorithmes de calcul rapide.

La définition de telles transformations repose sur les propriétés algébriques des entiers modulo M, pour certains choix de M ; on peut les désigner par transformations algébriques.

Le choix du module M est régi par les considérations suivantes :

a) Simplicité des calculs dans une arithmétique modulo.

Dans son principe, une arithmétique modulo implique une division par le module M. Cette division est triviale pour $M = 2^m$; elle est très simple pour $M = 2^m \pm 1$, car il suffit pour obtenir le résultat d'ajouter une retenue (arithmétique en complément à un) ou de la soustraire.

b) Le module doit être suffisamment grand.

Le résultat de la convolution doit pouvoir être représenté sans ambiguïté dans l'arithmétique modulo M. Par exemple, une convolution à 32 termes avec des données à 12 bits et des coefficients à 8 bits impose $M > 2^{25}$.

c) Propriétés algébriques convenables.

L'ensemble des entiers modulo M doit présenter des propriétés algébriques permettant de définir des transformations comparables à la TFD.

D'abord il doit exister dans cet ensemble des éléments périodiques, pour que l'on puisse élaborer des algorithmes de calcul rapide ; il faut disposer d'un élément α tel que :

$$\alpha^N = 1$$

On peut alors définir une transformation par l'expression :

$$X(k) = \sum_{n=0}^{N-1} x(n)\,\alpha^{nk} \qquad (3.45)$$

Pour que la Transformation inverse, définie par l'expression :

$$x(n) = N^{-1} \sum_{k=0}^{N-1} X(k)\,\alpha^{-nk} \qquad (3.46)$$

existe, il faut d'abord que N et les puissances de α possèdent des inverses.

On démontre que N a un inverse modulo M si N et M sont premiers entre eux. L'élément α doit être premier avec M et d'ordre N, c'est-à-dire que $\alpha^N = 1$.

L'existence de la transformation inverse implique de plus une condition sur les α^i, c'est-à-dire que :

$$\sum_{k=0}^{N-1} \alpha^{ik} = N\delta(i) \quad \text{avec} \quad \begin{array}{ll} \delta(i) = 1 & \text{si} \quad i = 0 \text{ modulo N} \\ \delta(i) = 0 & \text{si} \quad i \neq 0 \text{ modulo N} \end{array}$$

Cette condition se traduit par le fait que les éléments $(1 - \alpha^i)$ doivent posséder un inverse. On démontre que l'ensemble des conditions pour l'existence d'une transformation et de son inverse se ramène à la condition suivante : pour tout facteur premier P de M, il faut que N divise $P - 1$. Ainsi, si M est premier, N doit diviser $M - 1$.

Des algorithmes de calcul rapide peuvent être élaborés si N est un nombre composite, en particulier si N est une puissance de 2 ; ces algorithmes sont semblables à ceux de la TFR.

D'autre part les calculs à faire dans la transformation se trouvent considérablement simplifiés dans le cas particulier où $\alpha = 2$.

Finalement un choix du module M très intéressant est le suivant :

$$M = 2^{2m} + 1$$

quand M est premier. Ces nombres sont les nombres de Fermat.

Une transformation algébrique basée sur les nombres de Fermat est définie comme suit :

– Module $M = 2^{2m} + 1$
– Ordre de la Transformée : $N = 2^{m+1}$
– Transformée directe :

$$X(k) = \sum_{n=0}^{N-1} x(n) 2^{nk}$$

– Transformée inverse :

$$x(n) = (2^t) \sum_{k=0}^{N-1} X(k) 2^{-nk} \quad \text{avec} \quad t = 2^{m+1} - m - 1$$

Exemple : $m = 3$, $2^m = 8$, $M = 257$, $N = 16$, $t = 12$.

Cette transformation permet de calculer des convolutions de nombres réels, comme la Transformation de Fourier Discrète, mais avec les avantages suivants :

– Le résultat est obtenu sans approximation.
– Les opérations portent sur des nombres réels.
– Le calcul de la transformée et de son inverse ne nécessite aucune multiplication. Seules restent les multiplications dans l'espace transformé.

Cependant cette technique présente des limitations importantes. Les calculs étant exacts, le module M doit être suffisamment grand, ce qui conduit à des nombres de grande longueur.

Les relations entre les paramètres M et N données ci-dessus, imposent que les calculs soient faits avec un nombre de bits B de l'ordre de $\frac{N}{2}$; c'est-à-dire que le nombre de termes de la convolution est approximativement le double du nombre de bits des données dans le calcul. L'application se trouve par suite restreinte aux convolutions comportant peu de termes.

Le domaine d'application des transformations algébriques peut être élargi en faisant appel à d'autres nombres que les nombres de Fermat, ou encore en traitant les convolutions à grands nombres de termes comme des convolutions à deux dimensions [17]. La référence [18] décrit un exemple de réalisation.

Les transformations algébriques sont utilisées en codage correcteur d'erreur, notamment pour les Codes de Reed-Solomon.

BIBLIOGRAPHIE

[1] M. C. Pease – *Methods of Matrix Algebra.* Academic Press, New York, 1965.

[2] C. S. Burrus and T. W. Parks – *DFT/FFT and Convolution Algorithms*, Wiley, New York, 1985.

[3] H. Sloate – Matrix Representations for sorting and the Fast Fourier Transform. *IEEE Trans.* Vol. CAS-21, N° 1, Janvier 1974.

[4] G. Bergland – A fast Fourier Transform Algorithm for Real Valued Series. *Communications of th ACM,* Vol. 11, N° 10, Octobre 1968.

[5] J. L. Vernet – *Real Signals FFT by Means of an Odd Discrete Fourier Transform.* Proceedings of IEEE, Octobre 1971.

[6] G. Bonnerot and M. Bellanger – *Odd-Time Odd-Frequency DFT for Symmetric Real-Valued Series.* Proceedings of IEEE, March 1976.

[7] H. Ziegler – A fast Transform Algorithm for Symmetric Real Valued Series. *IEEE Transactions,* Vol. AU-20, N° 5, December 1972.

[8] C. Diab, M. Oueidat and R. Prost, « A New IDCT-DFT Relationship Reducing the IDCT computational Cost », IEEE Trans. on Signal Processing, Vol. 50, N° 7, July 2002, pp. 1681-84.

[9] J. Prado – « Transformation de Hartley Discrète Rapide », Annales des Télécom., 40, N°s 9-10, Oct. 1985, pp. 478-484.

[10] M. Vetterli, H. J. Nussbaumer – « Algorithmes de transformation de Fourier et en cosinus mono et bi-dimensionnels », Annales des Télécom., 40, N°s 9-10, Oct. 1985, pp. 466-477.

[11] M. A. Haque – « A two-Dimensionnal Fast Cosine Transform », *IEEE Trans.,* vol. ASSP-33, N° 6, Déc. 1985, pp. 1532-1539.

[12] H. S. Malvar, *Signal Processing with Lapped Transforms,* Artech House, Norwood MA, 1992.

[13] A. Despain – Very Fast Fourier Transform Algorithms Hardware for Implementation. *IEEE Transactions on Computers,* Vol. C. 28, N° 5, May 1979.

[14] H. Silverman – An Introduction to Programming the Winograd Fourier Transform Algorithm. *IEEE Transactions,* Vol. ASSP-25, N° 2, April 1977.

[15] H. Nussbaumer – Nouveaux Algorithmes de Transformée de Fourier Rapide. *L'onde Électrique,* Vol. 59, N°s 6-7, Juin 1979, pp. 81-88.

[16] B. Fino and R. Algazi, « Unified Matrix Treatment of Fast Walsh-Hadamard Transform », IEEE Transactions on Computers, Nov. 1976, pp. 1142-46.

[17] R. C. Agrawal and C. S. Burrus – *Number Theoretic Transforms to Implement Fast Digital Convolution.* Proc. IEEE, Vol. 63, 1975.

[18] J. H. Mac Clellan – Hardware Realization of a Fermat Number Transform. *IEEE Trans.* Vol. ASSP-24, June 1976.

EXERCICES

1 Effectuer le produit de Kronecker $A \otimes I_3$, de la matrice A telle que :

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

par la matrice unité I_3 de dimension 3.

Effectuer le produit $I_3 \otimes A$ et le comparer au précédent.

2 En prenant des matrices carrées de dimension 2, vérifier les propriétés des produits de Kronecker des matrices données au paragraphe 3.1.

3 Écrire la factorisation de la matrice de la TFD d'ordre 64 en base 2, en base 4 et en base 8, selon la procédure donnée au paragraphe 3.2.

4 En utilisant l'entrelacement temporel, factoriser la matrice de la TFD d'ordre 12. Quel est le nombre minimum de multiplications et additions nécessaire ? Écrire le programme de calcul.

5 Factoriser la matrice de la TFD d'ordre 16 en base 2, dans les deux cas suivants :

– les données se présentent en entrée et sortie dans l'ordre naturel,
– les étages de calcul sont identiques.

Donner, dans le dernier cas, un schéma de réalisation en utilisant comme mémoires, des registres à décalage.

6 Soit à calculer une Transformée de Fourier Discrète d'ordre 16 portant sur des données réelles.

Écrire l'algorithme qui utilise une TFD d'ordre 8, pour ce calcul.

Écrire l'algorithme basé sur la TFD impaire.

Comparer ces algorithmes et les nombres d'opérations.

7 Soit à calculer une TFD d'ordre 12, en utilisant une factorisation du type donné au paragraphe 3.6 et avec les permutations indiquées pour les données.

Évaluer le nombre d'opérations et le comparer aux valeurs trouvées dans l'exercice 4.

8 Pour effectuer la convolution circulaire des deux suites : $x = (2, -2, 1, 0)$ et $h = (1, 2, 0, 0)$, on utilise une transformée algébrique de module M = 17 et de coefficient $\alpha = 4$.

Comme N = 4, vérifier que $\alpha^N = 1$. Écrire la matrice de la transformation et de la transformation inverse. Vérifier que le résultat cherché est la suite $y = (2, 2, -3, 2)$.

Chapitre 4

Les systèmes linéaires discrets invariants dans le temps

Les systèmes linéaires discrets invariants dans le temps (LIT) constituent un domaine très important du traitement numérique du signal, qui est celui des filtres numériques à coefficients fixes. Ces systèmes se caractérisent par le fait que leur fonctionnement est régi par une équation de convolution. L'analyse de leurs propriétés se fait à l'aide de la Transformation en Z, qui joue pour les systèmes discrets le même rôle que la transformée de Laplace ou de Fourier pour les systèmes continus. Dans le présent chapitre les éléments les plus utiles pour l'étude de tels systèmes sont introduits brièvement. En complément on peut se reporter aux références [1, 2, 3, 4, 5].

4.1 DÉFINITION ET PROPRIÉTÉS

Un système discret est un système qui convertit une suite de données d'entrée $x(n)$ en une suite de sortie $y(n)$. Il est linéaire si la suite $x_1(n) + ax_2(n)$ est convertie en la suite $y_1(n) + ay_2(n)$. Il est invariant dans le temps si la suite $x(n - n_0)$ est convertie en la suite $y(n - n_0)$ quel que soit n_0 entier.

Soit $u_0(n)$ la suite unitaire représentée sur la figure 4.1. et définie par :

$$\begin{aligned} u_0(n) &= 1 \quad \text{pour} \quad n = 0 \\ u_0(n) &= 0 \quad \text{pour} \quad n \neq 0 \end{aligned} \tag{4.1}$$

Toute suite $x(n)$ peut se décomposer en une somme de suites unitaires convenablement décalées :

$$x(n) = \sum_{m=-\infty}^{\infty} x(m)u_0(n-m) \tag{4.2}$$

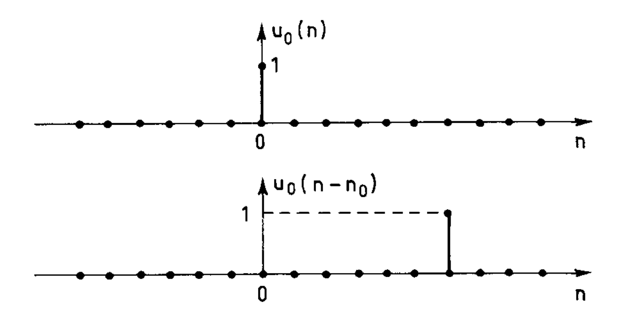


FIG. 4.1. *Suites unitaires*

D'autre part soit $h(n)$ la suite qui constitue la réponse du système à la suite unitaire $u_0(n)$. A la suite $u_0(n-m)$ correspond la réponse $h(n-m)$ en raison de l'invariance temporelle. La linéarité entraîne alors la relation suivante :

$$y(n) = \sum_m x(m)\,h(n-m) = \sum_m h(m)\,x(n-m) = h(n) * x(n) \quad (4.3)$$

C'est l'équation de convolution qui caractérise le système linéaire invariant dans le temps (LIT). Un tel système est donc complètement défini par la donnée de la suite $h(n)$, qui est appelée réponse impulsionnelle du système.

Ce système possède la propriété de causalité si la sortie à l'indice $n = n_0$ ne dépend que des entrées aux indices $n \leqslant n_0$. Cette propriété implique que $h(n) = 0$ pour $n < 0$, et la sortie est donnée par :

$$y(n) = \sum_{m=0}^{\infty} h(m)\,x(n-m) \quad (4.4)$$

Un système LIT est stable si à toute entrée d'amplitude bornée correspond une sortie bornée. Une condition nécessaire et suffisante de stabilité est donnée par l'inégalité :

$$\sum_n |h(n)| < \infty \quad (4.5)$$

Pour montrer que la condition est nécessaire il suffit d'appliquer au système la suite d'entrée $x(n)$ telle que :

$$x(n) = \begin{cases} +1 & \text{si} \quad h(n) \geqslant 0 \\ -1 & \text{si} \quad h(n) < 0 \end{cases}$$

Il vient alors pour $n = 0$:

$$y(0) = \sum_m |h(m)|$$

Si l'inégalité (4.5) n'est pas vérifiée, $y(0)$ n'est pas borné et le système n'est pas stable.

Si la suite d'entrée est bornée, c'est-à-dire :

$$|x(n)| \leqslant \mathrm{M} \quad \text{pour tout } n$$

alors il vient :

$$|y(n)| \leqslant \sum_m |h(m)|\,|x(n-m)| \leqslant \mathrm{M} \sum_m |h(m)|$$

Si l'égalité (4.5) est vérifiée, alors $y(n)$ est borné ; la condition est suffisante.

En particulier le système LIT défini par la réponse suivante :

$$h(m) = a^m \quad \text{avec} \quad m \geqslant 0$$

est stable pour $|a| < 1$.

Les caractéristiques des systèmes LIT sont étudiées à l'aide de la transformation en Z.

4.2 LA TRANSFORMATION EN Z

La transformée en Z, X(Z) de la suite $x(n)$ est définie par la relation suivante :

$$\mathrm{X(Z)} = \sum_{n=-\infty}^{\infty} x(n)\,\mathrm{Z}^{-n} \tag{4.6}$$

Z est une variable complexe et la fonction X(Z) possède un domaine de convergence qui en général est un anneau centré sur l'origine, de rayons R1 et R2. C'est-à-dire que X(Z) est défini pour $\mathrm{R1} < |\mathrm{Z}| < \mathrm{R2}$. Les valeurs R1 et R2 dépendent de la suite $x(n)$. Si la suite $x(n)$ représente la suite des échantillons d'un signal prélevés avec la période T, la transformée de Fourier de cette suite s'écrit :

$$\mathrm{S}(f) = \sum^{\infty} x(n) e^{-j2\pi f n\mathrm{T}}$$

Ainsi, pour $\mathrm{Z} = e^{j2\pi f\mathrm{T}}$ la transformée en Z de la suite $x(n)$ coïncide avec sa transformée de Fourier. C'est-à-dire que l'analyse d'un système discret peut se faire avec la transformée en Z et, pour connaître la réponse en fréquence, il suffit de remplacer Z par $e^{j2\pi f\mathrm{T}}$.

Cette transformée possède une transformée inverse. Soit Γ un contour fermé contenant tous les points singuliers, ou pôles, de X(Z) ainsi que l'origine ; on peut écrire :

$$\mathrm{Z}^{m-1}\,\mathrm{X(Z)} = \sum_{n=-\infty}^{\infty} x(n)\,\mathrm{Z}^{m-1-n} = x(m)\,\mathrm{Z}^{-1} + \sum_{n \neq m} \mathrm{Z}^{m-1-n}\,x(n)$$

et d'après le théorème des résidus :

$$x(m) = \frac{1}{2\pi j} \int_\Gamma \mathrm{Z}^{m-1}\,\mathrm{X(Z)}\,d\mathrm{Z} \tag{4.7}$$

Par exemple si $X(Z) = \dfrac{1}{1 - pZ^{-1}}$, on obtient par application directe de l'équation ci-dessus :

$$\begin{aligned} x(n) &= p^n \quad \text{pour} \quad n \geqslant 0 \\ x(n) &= 0 \quad \text{pour} \quad n < 0 \end{aligned}$$

De même à X(Z) définie par :

$$X(Z) = \sum_{i=1}^{N} \frac{a_i}{1 - p_i Z^{-1}}$$

correspond la suite $x(n)$ telle que :

$$\begin{aligned} x(n) &= \sum_{i=1}^{N} a_i p_i^n \quad \text{pour} \quad n \geqslant 0 \\ x(n) &= 0 \qquad\qquad \text{pour} \quad n < 0 \end{aligned}$$

Une condition de stabilité apparaît très simplement en observant que la suite $x(n)$ est bornée si, et seulement si $|p_i| < 1$ pour $1 \leqslant i \leqslant N$, c'est-à-dire que les pôles de X(Z) sont à l'intérieur du cercle unité.

Dans ces exemples, les termes de la suite $x(n)$ peuvent aussi être obtenus directement par développement en série. Quand X(Z) est une fraction rationnelle une méthode très simple pour obtenir les premières valeurs de la suite $x(n)$ consiste à faire une division de polynômes. Par exemple pour :

$$X(Z) = \frac{1 + 2Z^{-1} + Z^{-2} + Z^{-3}}{1 - Z^{-1} - 8Z^{-2} + 12Z^{-3}}$$

la division directe donne :

$$X(Z) = 1 + 3Z^{-1} + 12Z^{-2} + 25Z^{-3} + \ldots$$

d'où :

$$x(0) = 1\,; \quad x(1) = 3\,; \quad x(2) = 12\,; \quad x(3) = 25$$

La transformation en Z possède la propriété de linéarité. D'autre part la transformée en Z de la suite retardée $x(n - n_0)$ s'écrit :

$$X_{n_0}(Z) = Z^{-n_0} X(Z) \tag{4.8}$$

Ces deux propriétés sont utilisées pour calculer la transformée en Z, Y(Z), de la suite $y(n)$ obtenue en sortie d'un système linéaire discret, par convolution des suites $x(n)$ et $h(n)$ qui ont pour transformées X(Z) et H(Z).

En calculant la transformée en Z des deux membres de l'équation de convolution (4.3) :

$$y(n) = \sum_m h(m)\, x(n - m)$$

il vient :

$$Y(Z) = \sum_m h(m)\, Z^{-m}\, X(Z) = H(Z).X(Z) \tag{4.9}$$

Par suite la transformée en Z d'un produit de convolution est le produit des transformées. La fonction H(Z) est appelée fonction de transfert en Z du système LIT considéré.

La transformée en Z du produit de deux suites $x_3(n) = x_1(n).x_2(n)$ est la fonction $X_3(Z)$ définie par :

$$X_3(Z) = \frac{1}{2\pi j} \int_\Gamma X_1(v)\, X_2\left(\frac{Z}{v}\right) v^{-1}\, dv \tag{4.10}$$

Le contour d'intégration est à l'intérieur du domaine de convergence des fonctions $X_1(v)$ et $X_2\left(\frac{Z}{v}\right)$.

L'application aux suites causales amène à introduire la transformation en Z monolatérale.

La transformation en Z monolatérale de la suite $x(n)$ s'écrit :

$$X(Z) = \sum_{n=0}^{\infty} x(n)\, Z^{-n} \tag{4.11}$$

Les propriétés sont les mêmes que celles de la transformation définie par la relation (4.6), sauf pour les suites retardées. En effet la transformée de la suite $x(n - n_0)$ s'écrit :

$$X_{n_0}(Z) = \sum_{n=0}^{\infty} x(n-n_0)\, Z^{-n} = Z^{-n_0}.X(Z) + \sum_{n=1}^{n_0} x(-n)\, Z^{-(n_0-n)} \tag{4.12}$$

L'intérêt de cette transformation est de prendre en compte les conditions initiales et de faire apparaître les régimes transitoires dans l'étude de la réponse d'un système. D'autre part elle permet de déterminer à partir de X(Z) les valeurs extrêmes de la suite $x(n)$. La valeur initiale $x(0)$ s'écrit :

$$x(0) = \lim_{Z \to \infty} X(Z) \tag{4.13}$$

et la valeur finale, obtenue en calculant la transformée de la suite $x(n+1) - x(n)$:

$$x(\infty) = \lim_{Z \to 1} (Z-1)\, X(Z) = \lim_{z \to 1} (1 - z^{-1})X(z) \tag{4.14}$$

Pour des développements plus importants sur la transformation en Z et ses applications, on peut se reporter à la référence [6].

Les résultats ci-dessus s'appliquent au calcul de la puissance des signaux discrets.

4.3 ÉNERGIE ET PUISSANCE DES SIGNAUX DISCRETS

Soit à calculer l'énergie E du signal représenté par la suite $x(n)$, dont la Transformée en Z s'écrit X(Z). Par définition :

$$E = \sum_{n=-\infty}^{\infty} |x(n)|^2$$

La suite $x_3(n)$ définie par :

$$x_3(n) = |x(n)|^2$$

peut être considérée comme le produit de deux suites, $x_1(n)$ et $x_2(n)$ telles que :

$$x_1(n) = x(n); \quad x_2(n) = \bar{x}(n)$$

La transformée $X_3(Z)$ se calcule à partir des fonctions $X_1(Z)$ et $X_2(Z)$ à l'aide de la formule (4.10) donnée au paragraphe précédent pour la transformée en Z du produit de deux suites. L'évaluation au point Z = 1 conduit à la relation :

$$X_3(1) = \sum_{n=-\infty}^{\infty} |x(n)|^2 = \frac{1}{2\pi j} \int_\Gamma X_1(v)\, X_2\left(\frac{1}{v}\right) \frac{dv}{v}$$

Si Γ est le cercle unité, $\frac{1}{v} = \bar{v}$ et par suite :

$$X_2\left(\frac{1}{v}\right) = X_2(\bar{v}) = \sum_{n=-\infty}^{\infty} \bar{x}(n)\, (\bar{v})^{-n} = \overline{X}(v)$$

et comme alors $v = e^{j2\pi f}$, il vient :

$$E = \sum_{n=-\infty}^{\infty} |x(n)|^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} |X(e^{j2\pi f})|^2\, df \qquad (4.15)$$

C'est la relation de Bessel-Parseval donnée au paragraphe 1.1.1 qui exprime la conservation de l'énergie pour les signaux discrets : l'énergie du signal est égale à l'énergie contenue dans son spectre.

Les calculs ci-dessus font apparaître une expression utile pour la norme $\|X\|_2$ de la fonction $X(f)$; en effet, par définition :

$$\|X\|_2^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} |X(f)|^2\, df$$

Il vient :

$$\|X\|_2^2 = \frac{1}{2\pi j} \int_{|Z|=1} X(Z)\, X(Z^{-1})\, \frac{dZ}{Z} \qquad (4.16)$$

Si X(Z) est une fonction holomorphe de la variable complexe dans un domaine contenant le cercle unité, l'intégrale se calcule par la méthode des résidus et fournit directement la valeur de $\|X\|_2$ qui est aussi la norme L_2 du signal discret $x(n)$.

Soit maintenant à calculer l'énergie E_y du signal $y(n)$ en sortie du système LIT de réponse impulsionnelle $h(n)$ auquel est appliqué le signal $x(n)$.

Le signal $x(n)$ est d'abord supposé déterministe. D'après la relation (4.15), on peut écrire, en posant $\omega = 2\pi f$:

$$E_y = \frac{1}{2\pi} \int_{-\pi}^{\pi} |Y(e^{j\omega})|^2\, d\omega$$

La relation (4.9) fournit directement le résultat suivant :

$$E_y = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2\, |X(e^{j\omega})|^2 d\omega \qquad (4.17)$$

Ces résultats s'étendent aux signaux aléatoires.

4.4 FILTRAGE DES SIGNAUX ALÉATOIRES

Si le signal $x(n)$ est aléatoire et possède un moment d'ordre 1, $E[x(n)]$, on peut calculer l'espérance de la sortie $y(n)$ du système LIT. Il vient :

$$E[y(n)] = \sum_m h(m)\, E[x(n-m)] \qquad (4.18)$$

Si l'espérance de $x(n)$ est stationnaire, il en est de même de celle de $y(n)$, pourvu que le système soit stable, c'est-à-dire qu'il vérifie la relation (4.5).

Pour un signal $x(n)$ stationnaire d'ordre deux, de fonction d'autocorrélation $r_{xx}(n)$, on peut calculer la fonction d'autocorrélation $r_{yy}(n)$ de la sortie du système LIT. D'après l'expression de définition (1.58), on a :

$$r_{yy}(n) = E[y(i)\, y(i-n)] = \sum_m h(m) \,.\, E[x(i-m)\, y(i-n)]$$

En faisant apparaître la fonction de corrélation $r_{xy}(n)$ entre $x(n)$ et $y(n)$:

$$r_{xy}(n) = E[x(i)\, y(i-n)] \qquad (4.19)$$

il vient :

$$r_{yy}(n) = \sum_m h(m)\, r_{xy}(n-m) = h(n) * r_{xy}(n) \qquad (4.20)$$

Puis :

$$r_{xy}(n) = \sum_m h(m)\,\mathrm{E}[x(i)\,x(i-n-m)] = h(-n) * r_{xx}(n)$$

Finalement, on obtient le résultat suivant :

$$r_{yy}(n) = h(n) * h(-n) * r_{xx}(n) \qquad (4.21)$$

Alors, entre les transformées en Z, $\Phi_{xx}(z)$ et $\Phi_{yy}(\mathrm{Z})$ il existe la relation :

$$\Phi_{yy}(\mathrm{Z}) = \mathrm{H}(\mathrm{Z}).\mathrm{H}(\mathrm{Z}^{-1}).\Phi_{xx}(\mathrm{Z}) \qquad (4.22)$$

Cette expression peut constituer une approche plus commode que (4.21) pour calculer par transformation inverse la fonction d'autocorrélation du signal de sortie. Avec la transformée de Fourier, il vient :

$$r_{yy}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\mathrm{H}(e^{j\omega})|^2\,\Phi_{xx}(e^{j\omega})\,e^{jn\omega}\,d\omega \qquad (4.23)$$

En particulier, l'énergie du signal de sortie s'écrit :

$$\mathrm{E}_y = \varphi_{yy}(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\mathrm{H}(e^{j\omega})|^2\,\Phi_{xx}(e^{j\omega})\,d\omega \qquad (4.24)$$

C'est l'équivalent de la relation (4.17) pour les signaux aléatoires.

Il arrive que le signal $x(n)$ puisse être assimilé à un bruit blanc de variance σ_x^2. Alors la variance σ_y^2 du signal de sortie $y(n)$, est donnée par :

$$\sigma_y^2 = \frac{\sigma_x^2}{2\pi}\int_{-\pi}^{\pi} |\mathrm{H}(e^{j\omega})|^2\,d\omega \qquad (4.25)$$

ou encore en utilisant l'égalité (4.15) :

$$\sigma_y^2 = \sigma_x^2.\left[\sum_m h^2(n)\right] \qquad (4.26)$$

Ces résultats, d'un grand intérêt pratique, sont souvent utilisés par la suite, par exemple pour l'évaluation des puissances de bruit de calcul dans les filtres.

4.5 SYSTÈMES DÉFINIS PAR UNE ÉQUATION AUX DIFFÉRENCES

Les systèmes LIT, les plus intéressants sont les systèmes où les suites d'entrée et sortie sont liées par une équation aux différences linéaire à coefficients constants. En effet d'une part ils correspondent à des réalisations simples et d'autre part ils constituent une excellente modélisation de nombreux systèmes naturels.

Un système de ce type d'ordre N est défini par la relation suivante :

$$y(n) = \sum_{i=0}^{N} a_i x(n-i) - \sum_{i=1}^{N} b_i y(n-i) \tag{4.27}$$

En appliquant la transformation en Z aux deux membres de cette équation, et en désignant par Y(Z) et X(Z) les transformés des suites $y(n)$ et $x(n)$, on obtient :

$$Y(Z) = \sum_{i=0}^{N} a_i Z^{-i} X(Z) - \sum_{i=0}^{N} b_i Z^{-i} Y(Z) \tag{4.28}$$

soit :

$$Y(Z) = H(Z)\,X(Z)$$

avec :

$$H(Z) = \frac{a_0 + a_1 Z^{-1} + \ldots + a_N Z^{-N}}{1 + b_1 Z^{-1} + \ldots + b_N Z^{-N}} \tag{4.29}$$

La fonction de transfert du système H(Z) est une fraction rationnelle. Les a_i et b_i sont les coefficients du système ; certains coefficients peuvent être nuls, ce qui est le cas par exemple quand les deux sommations de l'expression (4.27) portent sur des nombres de termes différents. Pour faire apparaître la réponse en fréquence, il suffit de remplacer dans H(Z), la variable Z par $e^{j2\pi f}$.

La fonction H(Z) s'écrit sous forme d'un quotient de deux polynômes N(Z) et D(Z) de degré N et qui possèdent N racines Z_i et P_i respectivement avec $1 \leqslant i \leqslant N$.

En mettant en évidence ces racines, une autre expression de H(Z) apparaît :

$$H(Z) = \frac{N(Z)}{D(Z)} = a_0 \frac{\prod_{i=1}^{N} (1 - Z_i Z^{-1})}{\prod_{i=1}^{N} (1 - P_i Z^{-1})} \tag{4.30}$$

où a_0 est un facteur d'échelle ; on peut écrire :

$$H(Z) = a_0 \frac{\prod_{i=1}^{N} (Z - Z_i)}{\prod_{i=1}^{N} (Z - P_i)} \tag{4.31}$$

Dans le plan complexe, Z est l'affixe d'un point courant M, P_i et Z_i $(1 \leqslant i \leqslant N)$ sont les affixes des pôles et des zéros de la fonction H(Z). On peut écrire :

$$Z - Z_i = MZ_i \,.\, e^{j\theta_i} \quad \text{et} \quad Z - P_i = MP_i \,.\, e^{j\varphi_i}$$

et par suite la fonction de transfert s'exprime aussi par :

$$H(Z) = a_0 \prod_{i=1}^{N} \frac{MZ_i}{MP_i} e^{j \sum_{i=1}^{N} (\theta_i - \varphi_i)} \qquad (4.32)$$

Il en résulte une interprétation graphique dans le plan complexe. La réponse en fréquence du système est obtenue quand le point courant M parcourt le cercle unité. La figure 4.2 représente le cas d'un système d'ordre N = 2.

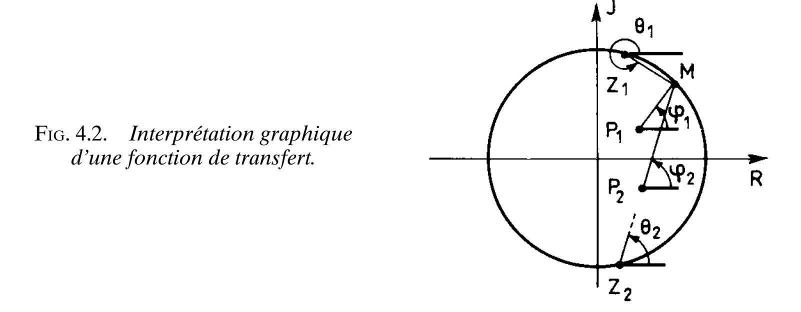


FIG. 4.2. *Interprétation graphique d'une fonction de transfert.*

Le module de la réponse en fréquence est ainsi égal au quotient du produit des distances du point courant M aux zéros Z_i par le produit des distances de M aux pôles P_i ; la phase est égale à la différence entre la somme des angles que font les vecteurs $\overrightarrow{P_iM}$ avec l'axe réel et la somme des angles que font les vecteurs $\overrightarrow{Z_iM}$ avec l'axe réel, pour suivre la convention prise au chapitre 1.

Cette interprétation graphique est très utilisée en pratique car elle offre une visualisation très simple de la forme de la réponse en fréquence d'un système.

En fait, l'analyse d'un système par sa réponse en fréquence correspond à un fonctionnement en régime permanent; elle est suffisante dans la mesure où les phénomènes transitoires peuvent être négligés. Si ce n'est pas le cas il faut introduire les conditions initiales, traduisant par exemple l'état d'un équipement et le contenu de ses mémoires à la mise sous tension.

Soit à étudier pour les valeurs de l'indice $n \geqslant 0$, le comportement du système défini par l'équation (4.27) auquel est appliqué la suite $x(n)$, nulle pour $n < 0$. La suite $y(n)$ est complètement déterminée si les valeurs $y(-i)$ avec $1 \leqslant i \leqslant N$ sont connues. Ces valeurs correspondent aux conditions initiales, et pour les introduire il faut faire appel à la transformation en Z monolatérale.

La transformation en Z monolatérale est appliquée aux deux membres de l'équation (4.27), en supposant que l'entrée $x(n)$ est un signal causal, c'est-à-dire que $x(n) = 0$ pour $n < 0$. Compte tenu de la relation (4.12) qui donne la transformée $Y_i(Z)$ de la suite retardée $y(n-i)$:

$$Y_i(Z) = Z^{-i} . Y(Z) + \sum_{n=1}^{i} y(-n)\, Z^{-(i-n)}$$

il vient :

$$Y(Z) = \sum_{i=0}^{N} a_i Z^{-i} X(Z) - \sum_{i=1}^{N} b_i Z^{-i} . Y(Z) - \sum_{i=1}^{N} b_i \sum_{n=1}^{i} y(-n) Z^{-(i-n)}$$

ou encore

$$Y(Z) = H(Z) X(Z) - \frac{\sum_{i=1}^{N} b_i \sum_{n=1}^{i} y(-n) Z^{-(i-n)}}{1 + \sum_{i=1}^{N} b_i Z^{-i}} \tag{4.33}$$

La réponse du système à l'indice n, $y(n)$, est obtenue par développement en série ou transformation inverse.

Il faut noter que les valeurs $y(-i)$ représentent l'état d'un système à la mise en fonctionnement seulement si ce système n'a en mémoire que des nombres de la suite de sortie. En fait il utilise souvent d'autres variables internes qui peuvent être introduites dans l'analyse en vue d'une généralisation et pour faire apparaître d'autres aspects touchant en particulier à la réalisation des matériels.

4.6 ANALYSE PAR LES VARIABLES D'ÉTAT

L'état d'un système d'ordre N à l'instant n est défini par un ensemble d'au moins N variables internes représentées par un vecteur $U(n)$ appelé vecteur d'état. Son fonctionnement est régi par les relations entre ce vecteur d'état et les signaux d'entrée et sortie. Le fonctionnement d'un système linéaire auquel est appliquée la suite d'entrée $x(n)$ et qui fournit la suite de sortie $y(n)$ est caractérisé en théorie des systèmes par le couple de relations suivantes, appelées équations d'état [7] :

$$\begin{aligned} U(n+1) &= A . U(n) + B . x(n) \\ y(n) &= C^t . U(n) + d . x(n) \end{aligned} \tag{4.34}$$

A est appelée matrice du système, B la commande, C le vecteur d'observation et d le coefficient de transition. La suite $x(n)$ est l'innovation et $y(n)$ l'observation. La justification de ces dénominations, qui ont leur origine en automatique, apparaît dans la suite, en particulier aux chapitres 7 et 13.

La matrice A est une matrice carrée de dimension N, B et C des vecteurs de dimension N.

L'état du système à l'instant n est obtenu à partir de l'état initial à l'instant zéro par l'équation :

$$U(n) = A^n . U(0) + \sum_{i=1}^{n} A^{n-i} B . x(i-1) \tag{4.35}$$

Par suite le comportement d'un tel système dépend des puissances successives de la matrice A.

La fonction de transfert en Z du système est obtenue en prenant la transformée en Z des équations d'état (4.34). Il vient :

$$\begin{aligned}(\mathrm{ZI} - \mathrm{A})\,\mathrm{U(Z)} &= \mathrm{BX(Z)}\\ \mathrm{Y(Z)} &= \mathrm{C}^t\mathrm{U(Z)} + d\mathrm{X(Z)}\end{aligned}$$

avec I matrice unité de dimension N ; par suite :

$$\mathrm{H(Z)} = \mathrm{C}^t(\mathrm{ZI} - \mathrm{A})^{-1}\,\mathrm{B} + d \tag{4.36}$$

Les pôles de la fonction de transfert ainsi obtenue sont les valeurs de Z qui annulent le déterminant de la matrice (ZI – A), c'est-à-dire les racines du polynôme caractéristique de A. Par conséquent, les pôles de la fonction de transfert du système sont les valeurs propres de la matrice A, qui doivent rester en module inférieures à l'unité pour que la stabilité soit assurée. Ce résultat est en concordance avec l'équation de fonctionnement du système (4.35) ; en effet, en diago-nalisant la matrice A, on voit que c'est la condition pour que le vecteur $\mathrm{U}(n) = \mathrm{A}^n.\mathrm{U}(0)$ tende vers zéro quand n tend vers l'infini, situation qui correspond à l'évolution libre du système à partir d'un état initial U(0).

L'examen de la fonction de transfert du système (4.36) montre par ailleurs que, quand un système est spécifié par la relation d'entrée-sortie, il existe une certaine latitude dans le choix des paramètres d'état ; en effet, seules les valeurs propres de la matrice A sont imposées, et la matrice du système peut être remplacée par une matrice semblable $\mathrm{A}' = \mathrm{M}^{-1}\,\mathrm{AM}$, où M est une matrice inversible, qui a les mêmes valeurs propres. Alors pour conserver la même suite de sortie, d'après (4.35) il faut :

$$\mathrm{A}' = \mathrm{M}^{-1}\,\mathrm{AM}\,;\quad \mathrm{C}'^t = \mathrm{C}^t.\mathrm{M}\,;\quad \mathrm{B}' = \mathrm{M}^{-1}\mathrm{B}.$$

La matrice A peut aussi être remplacée par sa transposée A^t ; alors le système est décrit par un système d'équations dual de (4.34), correspondant au vecteur d'état $\mathrm{V}(n)$ tel que :

$$\begin{aligned}\mathrm{V}(n+1) &= \mathrm{A}^t.\mathrm{V}(n) + \mathrm{C}.x(n)\\ y(n) &= \mathrm{B}^t.\mathrm{V}(n) + d.x(n)\end{aligned} \tag{4.37}$$

Cette représentation d'état fournit un autre mode de réalisation du système.

Les résultats obtenus dans ce paragraphe sont utilisés par la suite pour étudier certaines propriétés et pour faire apparaître des structures de réalisation de systèmes LIT.

Une étude approfondie va être faite dans les chapitres suivants pour deux types de systèmes LIT définis par une équation aux différences, les filtres numériques à réponse impulsionnelle finie et infinie.

BIBLIOGRAPHIE

[1] L. R. Rabiner and B. Gold – *Theory and Application of Digital Signal Processing.* Chapitre II. Prentice Hall, 1975.

[2] A. V. Oppenheim and R. W. Schafer – *Digital Signal Processing.* Chapitre II. Prentice Hall, 1974.

[3] J. Liferman – *Les systèmes discrets*. Masson Éd. 1975.

[4] R. Boite et H. Leich – *Les filtres Numériques. Analyse et synthèse des filtres unidimensionnels,* Collection CNET-ENST, Éditions Masson, 1980.

[5] J. Max et collaborateurs – *Méthodes et Techniques de Traitement du Signal,* Éditions Masson, 1981.

[6] E. I. Jury – *Theory and Application of the Z-Transform Method.* John Wiley, 1964.

[7] J. E. Cadzow – *Discrete Time Systems.* Prentice-Hall, 1973.

EXERCICES

1 Soit un système LIT dont la réponse impulsionnelle $h(n)$ est telle que :

$$h(n) = 1, \quad 0 \leqslant n \leqslant 3$$
$$h(n) = 0, \quad n < 0 \quad \text{et} \quad n > 3.$$

Calculer la réponse $y(n)$ à la suite $x(n)$ telle que :

$$x(n) = a^n, \quad \text{avec} \quad a = 0{,}7 \quad \text{pour} \quad 0 \leqslant n \leqslant 5$$
$$x(n) = 0 \quad \text{ailleurs.}$$

Réponse à la suite :

$$x(n) = \cos\left(\frac{2\pi n}{8}\right) \quad \text{pour} \quad 0 \leqslant n \leqslant 7$$

$$x(n) = 0 \quad \text{ailleurs.}$$

2 Montrer que la Transformée en Z de la suite causale $x(n)$ définie par :

$$x(n) = n\mathrm{T}e^{-an\mathrm{T}} \quad \text{pour} \quad n \geqslant 0$$
$$x(n) = 0 \quad \text{pour} \quad n < 0$$

a pour expression :

$$\mathrm{X(Z)} = \frac{\mathrm{T}e^{-a\mathrm{T}}\,\mathrm{Z}^{-1}}{(1 - e^{-a\mathrm{T}}\,\mathrm{Z}^{-1})^2}$$

Calculer les transformées inverses de ln $(\mathrm{Z} - a)$, $\dfrac{\mathrm{Z}}{(\mathrm{Z}-a)(\mathrm{Z}-b)}$ et établir les conditions sur a et b pour que la suite obtenue converge.

3 Calculer la Transformée en Z de la réponse impulsionnelle

$$h(n) = r^n \frac{\sin[(n+1)\theta]}{\sin(\theta)}, \quad n \geqslant 0$$
$$h(n) = 0 \qquad n < 0$$

Quel est le domaine de convergence de la fonction obtenue ? Placer ses pôles et zéros dans le plan des Z.

4 Soit un système LIT dont la fonction de transfert H(Z) s'écrit :

$$\mathrm{H(Z)} = \frac{1}{1 - 1{,}6\,\mathrm{Z}^{-1} + 0{,}92\,\mathrm{Z}^{-2}}$$

et auquel est appliqué un signal à spectre uniforme et de puissance unité. Calculer la puissance du signal en sortie du système et donner la répartition spectrale.

5 Utiliser la Transformation en Z monolatérale pour calculer la réponse du système défini par l'équation aux différences :

$$y(n) = x(n) + y(n-1) - 0{,}8\,y(n-2)$$

avec les conditions initiales $y(-1) = a$ et $y(-2) = b$ à la suite $x(n)$ définie par :

$$x(n) = e^{jn\omega} \quad \text{pour} \quad n \geqslant 0$$
$$x(n) = 0 \quad \text{pour} \quad n < 0$$

Mettre en évidence la réponse due aux conditions initiales et la réponse en régime permanent.

Chapitre 5

Les filtres à réponse impulsionnelle finie (RIF)

Les filtres numériques à réponse impulsionnelle finie (RIF) sont des systèmes linéaires discrets invariants dans le temps définis par une équation selon laquelle un nombre de sortie, représentant un échantillon du signal filtré, est obtenu par sommation pondérée d'un ensemble fini de nombres d'entrée, représentant les échantillons du signal à filtrer. Les coefficients de la sommation pondérée constituent la réponse impulsionnelle du filtre et un ensemble fini d'entre eux seulement prennent des valeurs non nulles. Ce filtre est du type « à mémoire finie », c'est-à-dire qu'il détermine sa sortie en fonction d'informations d'entrée d'ancienneté limitée. Il est fréquemment désigné par filtre non récursif, en raison de sa structure, car il ne nécessite pas de boucle de réaction dans sa réalisation, comme c'est le cas pour une autre catégorie de filtres, celle des filtres à réponse impulsionnelle infinie.

Les propriétés des filtres RIF vont être mises en évidence sur deux exemples simples.

5.1 PRÉSENTATION DES FILTRES RIF

Soit un signal $x(t)$ représenté par ses échantillons $x(nT)$, prélevés à la fréquence $f_e = \frac{1}{T}$, et soit à déterminer l'incidence sur le spectre de ce signal de l'opération qui consiste à remplacer la suite $x(nT)$ par la suite $y(nT)$ définie par la relation :

$$y(nT) = 1/2\,[x(nT) + x((n-1)\,T)] \qquad (5.1)$$

Cette suite est aussi celle qui est obtenue par échantillonnage du signal $y(t)$ tel que :

$$y(t) = 1/2\,[x(t) + x(t-T)]$$

Si $Y(f)$ et $X(f)$ désignent les transformées de Fourier des signaux $y(t)$ et $x(t)$, il vient :

$$Y(f) = 1/2\, X(f)\, (1 + e^{-j2\pi f T})$$

L'opération étudiée correspond à la fonction de transfert

$$H(f) = Y(f)/X(f)$$

telle que :

$$H(f) = e^{-j\pi f T} \cos(\pi f T) \qquad (5.2)$$

C'est une opération de filtrage appelée filtrage en cosinusoïde, qui conserve la composante continue et élimine la composante à la fréquence $f_e/2$, comme il est aisé de le vérifier directement.

Dans l'expression de $H(f)$ le terme complexe $e^{-j\pi f T}$ caractérise un retard $\tau = \frac{T}{2}$ qui est le temps de propagation du signal à travers le filtre.

La réponse impulsionnelle $h(t)$ qui correspond au filtre de fonction de transfert $|H(f)|$ s'écrit :

$$h(t) = 1/2 \left[\delta\left(t + \frac{T}{2}\right) + \delta\left(t - \frac{T}{2}\right)\right]$$

La figure 5.1 représente les caractéristiques du filtre.

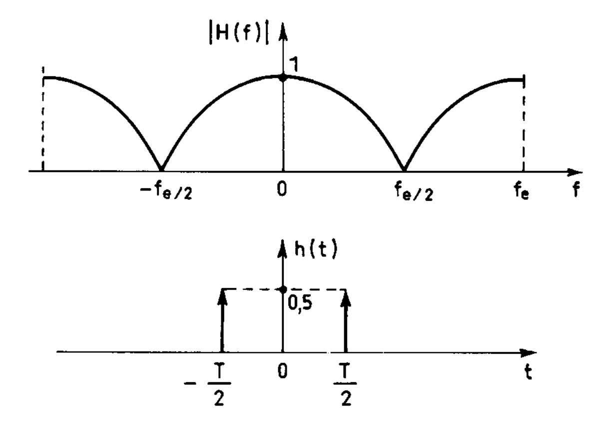


FIG. 5.1. *Le filtrage en cosinusoïde*

Une autre opération simple est celle qui associe à la suite des $x(nT)$ la suite des $y(nT)$ définie par :

$$y(nT) = 1/4\, [x(nT) + 2x\,[(n-1)T] + x\,[(n-2)T]] \qquad (5.3)$$

Comme la précédente, elle conserve la composante à la fréquence zéro et élimine celle à $f_e/2$. Elle correspond à la fonction de transfert :

$$H(f) = 1/4\,(1 + 2e^{-j2\pi f2T} + e^{-j2\pi f2T}) = e^{-j2\pi fT}\,1/2\,(1 + \cos 2\pi f T) \qquad (5.4)$$

Le filtre obtenu est dit en cosinusoïde surélevée ; son temps de propagation est $\tau = T$; à $|H(f)|$ correspond la réponse impulsionnelle $h(t)$ telle que :

$$h(t) = 1/4\,\delta(t + T) + 1/2\,\delta(t) + 1/4\,\delta(t - T).$$

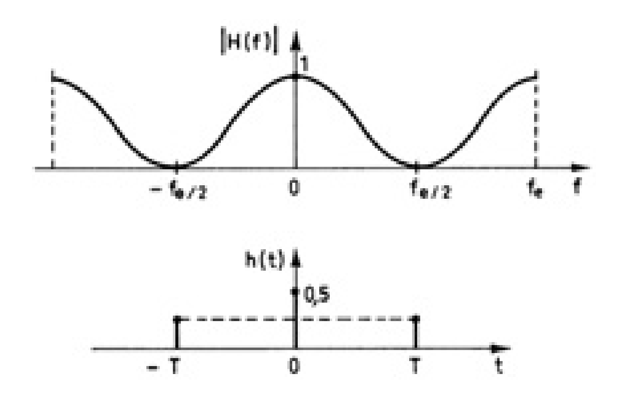


FIG. 5.2. *Filtre en cosinusoïde surélevée*

Ce filtre est un passe-bas plus sélectif que le précédent et il apparaît clairement que pour obtenir une fonction de filtrage plus sélective encore il suffit d'augmenter le nombre de termes de la suite $x(nT)$ sur lesquels porte la sommation pondérée.

Ces deux exemples ont permis de faire apparaître les caractéristiques suivantes des filtres RIF :

– La suite d'entrée $x(n)$ et la suite de sortie $y(n)$ sont reliées par une équation du type suivant qui constitue la relation de définition :

$$y(n) = \sum_{i=0}^{N-1} a_i x(n - i) \qquad (5.5)$$

Le filtre ainsi défini comporte un nombre N fini de coefficients a_i ; considéré comme un système discret, il a pour réponse à la suite unitaire la suite $h(i)$ telle que :

$$h(i) = a_i \quad \text{si} \quad 0 \leqslant i \leqslant N - 1$$
$$h(i) = 0 \qquad \text{ailleurs.}$$

C'est-à-dire que la réponse impulsionnelle est simplement la suite des coefficients.

– La fonction de transfert du filtre s'écrit :

$$\mathrm{H}(f) = \sum_{i=0}^{N-1} a_i e^{-j2\pi f i \mathrm{T}} \tag{5.6}$$

ou encore, exprimée en fonction de la variable Z :

$$\mathrm{H}(Z) = \sum_{i=0}^{N-1} a_i Z^{-i} \tag{5.7}$$

– La fonction $\mathrm{H}(f)$, réponse en fréquence du filtre, est une fonction périodique, de période $f_e = \frac{1}{\mathrm{T}}$. Les coefficients $a_i (0 \leqslant i \leqslant \mathrm{N} - 1)$ constituent le développement en série de Fourier de cette fonction.

La relation de Bessel-Parseval énoncée au paragraphe I.1.1 permet d'écrire :

$$\sum_{i=0}^{N-1} |a_i|^2 = \frac{1}{f_e} \int_0^{f_e} |\mathrm{H}(f)|^2 \, df \tag{5.8}$$

– Si les coefficients sont symétriques, la fonction de transfert peut se mettre sous la forme d'un produit de deux termes dont l'un est une fonction réelle et l'autre un nombre complexe de module 1 représentant un temps de propagation τ constant et égal à un multiple entier de la demi-période d'échantillonnage. Un tel filtre est dit à phase linéaire.

5.2 FONCTIONS DE TRANSFERT RÉALISABLES ET FILTRES À PHASE LINÉAIRE

Un filtre numérique traitant des nombres qui représentent les échantillons du signal prélevés avec la période T, a une réponse en fréquence périodique et de période $f_e = \frac{1}{\mathrm{T}}$. Par suite cette fonction $\mathrm{H}(f)$ est développable en série de Fourier :

$$\mathrm{H}(f) = \sum_{n=-\infty}^{\infty} \alpha_n \, e^{+j2\pi f n \mathrm{T}} \tag{5.9}$$

avec :

$$\alpha_n = \frac{1}{f_e} \int_0^{f_e} \mathrm{H}(f) \, e^{-j2\pi f n \mathrm{T}} \, df \tag{5.10}$$

Les coefficients α_n du développement sont à une constante près les échantillons prélevés avec la période T de la transformée de Fourier de la fonction $\mathrm{H}(f)$,

prise sur un intervalle de fréquence de largeur f_e. Comme ils constituent la réponse impulsionnelle, la condition de stabilité du filtre donnée par la relation (4.5) implique que les α_n tendent vers zéro quand n tend vers l'infini. Par suite la fonction H(f), peut être approchée par un développement limité à un nombre fini de termes :

$$\mathrm{H}(f) \simeq \sum_{n=-\mathrm{K}}^{\mathrm{L}} \alpha_n \, e^{j2\pi f n \mathrm{T}} = \mathrm{H_L}(f)$$

où K et L sont des entiers finis; l'approximation est d'autant meilleure que ces nombres sont plus grands.

La propriété de causalité, qui traduit le fait que dans un filtre réel la sortie ne peut précéder l'entrée dans le temps, implique que la réponse impulsionnelle $h(n)$ soit nulle pour $n < 0$. D'après les relations (5.5) et (5.6), si le filtre est causal, alors L = 0 et il vient :

$$\mathrm{H_L}(f) = \sum_{n=0}^{\mathrm{K}} a_n \, e^{-j2\pi f n \mathrm{T}}$$

Il en résulte que toute fonction de filtrage numérique stable et causale peut être approchée par la fonction de transfert d'un filtre RIF.

Le filtrage à phase linéaire correspond, pour la réponse en fréquence, à l'expression suivante :

$$\mathrm{H}(f) = \mathrm{R}(f) \, e^{-j\varphi(f)} \tag{5.11}$$

où $\mathrm{R}(f)$ est une fonction réelle où la phase $\varphi(f)$ est une fonction linéaire : $\varphi(f) = \varphi_0 + 2\pi f \tau$; τ est une constante donnant le temps de propagation à travers le filtre.

Il faut bien noter que cette condition ne correspond pas, en toute rigueur, à une linéarité de la phase. En effet, les changements de signe de $\mathrm{R}(f)$ amènent des discontinuités de π sur la phase; celle-ci peut se décomposer en une composante discrète et une composante continue, à laquelle la condition ci-dessus impose la linéarité. Cependant, par extension, les filtres étudiés sont dits à phase linéaire.

La réponse impulsionnelle d'un tel filtre s'écrit :

$$h(t) = e^{-j\varphi_0} \int_{-\infty}^{\infty} \mathrm{R}(f) \, e^{j2\pi f(t-\tau)} \, df \tag{5.12}$$

En supposant d'abord φ_0 nul et en décomposant la fonction réelle $\mathrm{R}(f)$ en une partie paire $\mathrm{P}(f)$ et une partie impaire $\mathrm{I}(f)$, il vient :

$$h(t+\tau) = 2 \int_0^{\infty} \mathrm{P}(f) \cos(2\pi f t) \, df + 2j \int_0^{\infty} \mathrm{I}(f) \sin(2\pi f t) \, df$$

Si l'on impose à la fonction $h(t)$ d'être réelle, il vient :

$$h(t+\tau) = 2 \int_0^{\infty} \mathrm{P}(f) \cos(2\pi f t) \, df$$

Cette relation montre que la réponse impulsionnelle est symétrique par rapport au point $t = \tau$ de l'axe des temps, c'est-à-dire que les coefficients du filtre doivent être symétriques. Deux configurations se présentent alors, suivant que le nombre de coefficients N est pair ou impair.

• $N = 2P + 1$: le filtre a un temps de propagation $\tau = PT$. La fonction de transfert s'écrit :

$$H(f) = e^{-j2\pi fPT}\left[h_0 + 2\sum_{i=1}^{P} h_i \cos(2\pi fiT)\right] \qquad (5.13)$$

• $N = 2P$: le filtre a comme temps de propagation $\tau = \left(P - \frac{1}{2}\right)T$. La fonction de transfert s'écrit :

$$H(f) = e^{-j2\pi f\left(P-\frac{1}{2}\right)T}\, 2\left(\sum_{i=1}^{P} h_i \cos\left(2\pi f\left(i - \frac{1}{2}\right)T\right]\right) \qquad (5.14)$$

Les h_i, coefficients du filtre, constituent la réponse du filtre numérique à la suite unitaire. En négligeant les repliements de spectre, ils peuvent aussi être considérés comme les échantillons, prélevés avec la période T, de la réponse impulsionnelle continue $h(t)$ du filtre qui a la même réponse en fréquence que le filtre numérique dans l'intervalle $\left(-\frac{1}{2T}, \frac{1}{2T}\right)$, mais sans la périodicité sur l'axe des fréquences. L'illustration est fournie par les figures 5.3 et 5.4 pour les cas où N est impair et pair, respectivement.

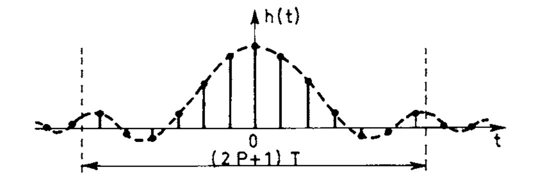


Fig. 5.3. *Filtre symétrique avec N impair*

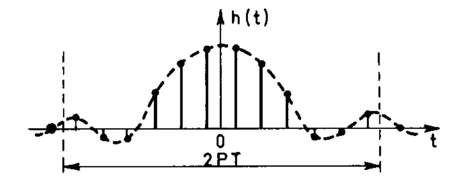


Fig. 5.4. *Filtre symétrique avec N pair*

Ces filtres font intervenir dans leur réponse en fréquence la fonction paire $P(f)$. Avec des coefficients réels on peut aussi obtenir une réponse en fréquence qui corresponde à la partie impaire de $R(f)$, la fonction $I(f)$.

Comme la fonction $h(t)$ doit être réelle, cette catégorie de filtres a pour fonction de transfert :

$$H(f) = -je^{-j2\pi f\tau}\, I(f) = e^{-j\frac{\pi}{2}}\, e^{-j2\pi f\tau}\, I(f)$$

Au déphasage proportionnel à la fréquence s'ajoute un déphasage fixe $\varphi_0 = \frac{\pi}{2}$ qui fait correspondre à un signal, le signal en quadrature. Cette possibilité

est intéressante dans certains types de modulation et est examinée ultérieurement. La réponse impulsionnelle est nulle au point $t = \tau$ et antisymétrique par rapport à ce point de l'axe des temps. Les configurations, suivant que le nombre de coefficients N est impair ou pair, sont représentées sur les figures 5.5 et 5.6 respectivement.

- N = 2P + 1 : le filtre a un temps de propagation $\tau = PT$

$$H(f) = -je^{-j2\pi f\tau}\, 2 \sum_{i=1}^{P} h_i \sin(2\pi f i T) \tag{5.15}$$

- N = 2P : le filtre a un temps de propagation $\tau = \left(P - \frac{1}{2}\right)T$

$$H(f) = -je^{-j2\pi f\tau}\, 2 \sum_{i=1}^{P} h_i \sin\left[2\pi f\left(i - \frac{1}{2}\right)T\right] \tag{5.16}$$

Comme $h_0 = 0$, la fonction de transfert a la même expression dans les deux cas.

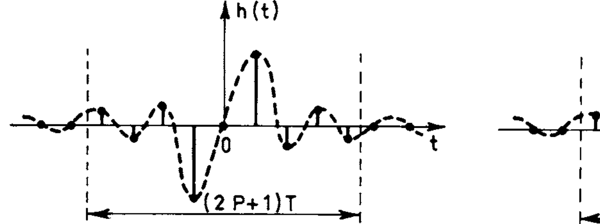


Fig. 5.5. *Filtre antisymétrique N impair*

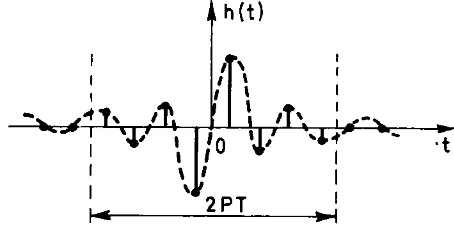


Fig. 5.6. *Filtre antisymétrique N pair*

Il est aisé de concevoir que des déphasages fixes, autres que $\varphi_0 = 0$ et $\varphi_0 = \frac{\pi}{2}$ peuvent être obtenus avec des filtres à coefficients complexes.

Le calcul des coefficients des filtres RIF va d'abord être étudié avec l'hypothèse de la phase linéaire, qui correspond à l'essentiel des applications et lorsque les spécifications sont données sur la réponse en fréquence.

5.3 CALCUL DES COEFFICIENTS PAR DÉVELOPPEMENT EN SÉRIE DE FOURIER POUR DES SPÉCIFICATIONS EN FRÉQUENCE

Les spécifications en fréquence correspondent à la donnée d'un gabarit.

Pour un filtre passe-bas on impose par exemple à la valeur absolue de la fonction de transfert d'approcher la valeur 1 avec la précision δ_1, dans la bande de fréquence $(0, f_1)$ dite bande passante et la valeur 0 avec la précision δ_2, dans la bande

$\left(f_2, \frac{f_e}{2}\right)$, dite bande affaiblie. Le gabarit correspondant est représenté sur la figure 5.7. L'intervalle $\Delta f = f_2 - f_1$ est appelé bande de transition et la raideur de coupure désigne le paramètre R_c tel que :

$$R_c = \frac{f_1 + f_2}{2(f_2 - f_1)} \tag{5.17}$$

FIG. 5.7. *Gabarit de filtre passe-bas*

Une méthode très simple pour obtenir les coefficients h_i consiste à développer en série de Fourier la fonction périodique H(f) à approcher ; il vient alors :

$$h_i = \frac{1}{f_e} \int_0^{f_e} H(f)\, e^{-j2\pi i \frac{f}{f_e}}\, df$$

Dans le cas du filtre passe-bas correspondant au gabarit de la figure 5.7, la relation (I.5) conduit à :

$$h_i = \frac{f_1 + f_2}{f_e} \cdot \frac{\sin \pi i \dfrac{f_1 + f_2}{f_e}}{\pi i \dfrac{f_1 + f_2}{f_e}} \tag{5.18}$$

Le tableau donné en annexe I du chapitre I peut ainsi être utilisé pour fournir une première estimation des valeurs des coefficients d'un filtre RIF dont, en fait, les valeurs optimisées calculées par la suite, s'écartent généralement assez peu.

Pour que le filtre soit réalisable il faut limiter à N le nombre de coefficients. Cette opération revient à multiplier la réponse impulsionnelle $h(t)$ par une fenêtre temporelle $g(t)$ telle que :

$$g(t) = 1 \quad \text{pour} \quad \frac{-NT}{2} \leqslant t \leqslant \frac{NT}{2}$$

$$g(t) = 0 \quad \text{ailleurs.}$$

La transformée de Fourier de cette fonction s'écrit en appliquant (1.10) :

$$G(f) = NT \frac{\sin(\pi f NT)}{\pi f NT} \quad (5.19)$$

La figure 5.8 montre ces fonctions.

Le filtre réel, à nombre limité N de coefficients, a pour fonction de transfert $H_R(f)$, le produit de convolution suivant :

$$H_R(f) = \int_{-\infty}^{\infty} H(f')\, G(f-f')\, df'$$

La limitation du nombre de coefficients introduit des ondulations et limite la raideur de coupure du filtre comme le montre la figure 5.9, qui correspond au cas où le filtre à réaliser est un passe-bas idéal de fréquence de coupure f_c.

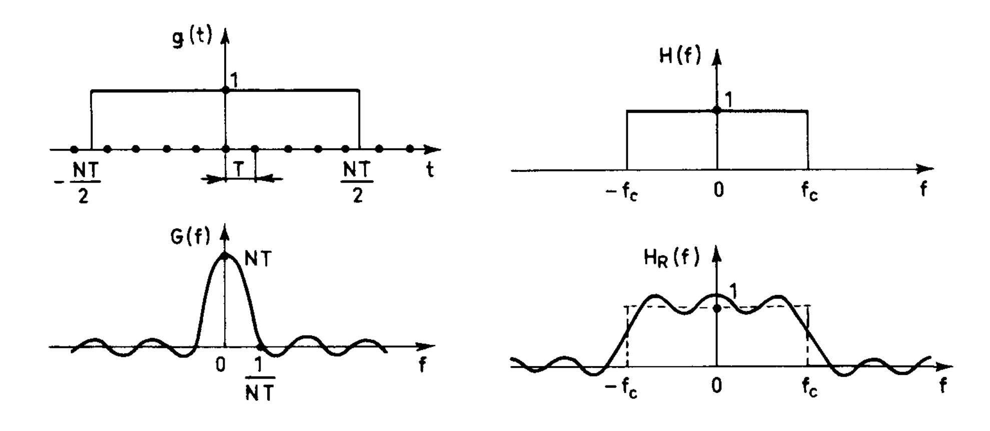


Fig. 5.8. *Fenêtre rectangulaire*

Fig. 5.9. *Incidence de la limitation du nombre de coefficients*

Les ondulations dépendent de celles de la fonction G(f) et, pour les réduire, il suffit de choisir comme fenêtre temporelle une fonction dont le spectre présente moins d'ondulations que celui de la fenêtre rectangulaire ci-dessus. C'est la situation exposée au paragraphe 2.4.2 pour l'analyse spectrale et on peut utiliser les mêmes fonctions, par exemple la fenêtre de Hamming définie comme suit :

$$g(t) = 0{,}54 + 0{,}46 \cos(2\pi t/NT) \quad \text{pour} \quad |t| \leqslant NT/2$$

$$g(t) = 0 \quad \text{pour} \quad |t| > NT/2$$

La contrepartie de la réduction des ondulations en bandes passante et affaiblie est un élargissement de la bande de transition.

La fonction qui présente les ondulations les plus faibles pour une largeur donnée du lobe principal, est la fonction dite de Dolf-Tchebycheff :

$$G(x) = \frac{\cos\,[K \cos^{-1}(Z_0 \cos \pi x)]}{\mathrm{ch}\,[K\,\mathrm{ch}^{-1}(Z_0)]} \quad \text{pour} \quad x_0 \leqslant x \leqslant 1 - x_0$$

$$G(x) = \frac{\mathrm{ch}\,[K\,\mathrm{ch}^{-1}(Z_0 \cos \pi x)]}{\mathrm{ch}\,[K\,\mathrm{ch}^{-1}(Z_0)]} \quad \text{pour} \quad 0 \leqslant x \leqslant x_0 \quad \text{et} \quad 1 - x_0 \leqslant x \leqslant 1 \tag{5.20}$$

avec $x_0 = \frac{1}{\pi} \cos^{-1}\left(\frac{1}{Z_0}\right)$; K est un nombre entier et Z_0 un paramètre. Cette fonction, que montre la figure 5.10, présente un lobe principal de largeur B, tel que :

$$B = 2 . x_0 = \frac{2}{\pi} \cos^{-1}\left(\frac{1}{Z_0}\right)$$

et des lobes secondaires d'amplitude constante égale à :

$$A = \frac{1}{\mathrm{ch}\,[K\,\mathrm{ch}^{-1}(Z_0)]}.$$

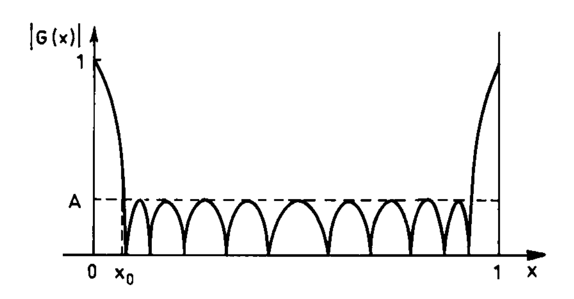


FIG. 5.10. *Fonction de Dolf-Tchebycheff*

Elle est périodique et sa transformée de Fourier inverse est constituée d'un ensemble de K + 1 valeurs discrètes non nulles, utilisées pour pondérer les coefficients du développement en série de Fourier de la fonction de filtrage à approcher.

Exemple

Soit à calculer les coefficients d'un filtre passe-bas de fréquence d'échantillonnage $f_e = 1$, fréquence de coupure $f_c = 0{,}25$, bande de transition $\Delta f = 0{,}115$ et comportant N = 17 coefficients.

La fonction de Dolf-Tchebycheff correspondante a pour paramètres K = 16 et Z_0 tel que :

$$2x_0 = \frac{2}{\pi} \cos^{-1}\left(\frac{1}{Z_0}\right) \simeq \Delta f.$$

Cette valeur correspond à des ondulations d'amplitude $A = \dfrac{1}{\text{ch}\,[16\ \text{ch}^{-1}\,(Z_0)]}$ dont la valeur a été prise à A = 0,1. La transformée de Fourier inverse $g(t)$ de cette fonction se compose de 17 valeurs discrètes non nulles qui sont données sur la figure 5.11, à un facteur d'échelle près.

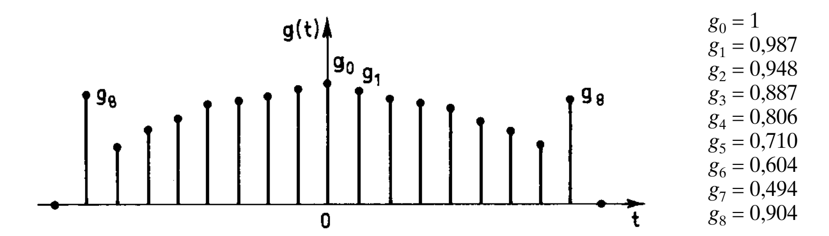


FIG. 5.11. *Coefficients de pondération d'une fenêtre de Dolf-Tchebycheff*

Ces valeurs constituent les coefficients de pondération de la réponse impulsionnelle $h(t)$ du filtre passe-bas idéal, donné à la figure 5.12.

$h_0 = 0{,}5$
$h_1 = 0{,}318$
$h_2 = 0$
$h_3 = -0{,}106$
$h_4 = 0$
$h_5 = 0{,}064$
$h_6 = 0$
$h_7 = -0{,}045$
$h_8 = 0$

FIG. 5.12. *Réponse impulsionnelle du filtre idéal*

Le filtre obtenu a pour coefficients les valeurs $a_i = g_i . h_i$ soit :

$a_0 = 0{,}5$	$a_4 = 0$
$a_1 = 0{,}3141$	$a_5 = 0{,}0451$
$a_2 = 0$	$a_6 = 0$
$a_3 = -0{,}0941$	$a_7 = -0{,}0224$
	$a_8 = 0$

La fonction de transfert est donnée par la figure 5.13.

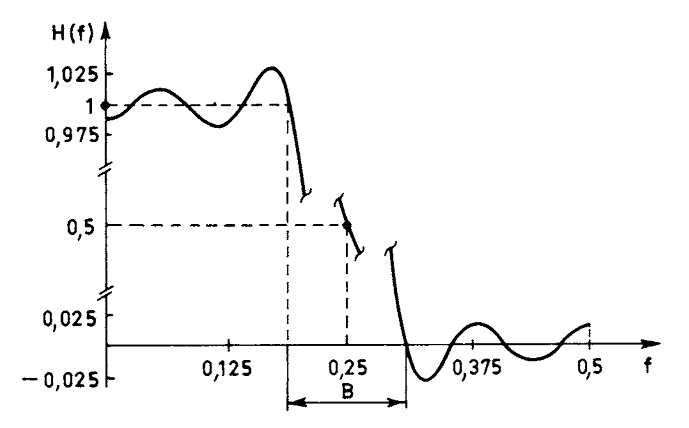


FIG. 5.13. *Fonction de transfert du filtre réel*

Il convient de remarquer que si les ondulations de la fonction $G(x)$ sont d'amplitude constante il n'en est plus de même des ondulations du filtre obtenu qui décroissent en amplitude quand on s'éloigne de la bande de transition. D'autre part les ondulations en bandes passante et affaiblie, sont les mêmes.

Ainsi, la technique du développement en série de Fourier de la fonction à approcher conduit à une détermination simple des coefficients du filtre mais elle implique deux limitations importantes :

– les ondulations du filtre sont égales en bandes passante et affaiblie,
– l'amplitude des ondulations n'est pas constante.

La première des limitations peut être levée par une méthode qui garde la simplicité du calcul direct, la méthode des moindres carrés. De plus, elle correspond précisément aux objectifs à atteindre dans un certain nombre d'applications.

5.4 CALCUL DES COEFFICIENTS PAR LA MÉTHODE DES MOINDRES CARRÉS

Soit à calculer les N coefficients h_i d'un filtre RIF de manière à ce que la fonction de transfert approche une fonction donnée suivant un critère des moindres carrés.

Le calcul peut se faire directement à partir de la relation entre les coefficients et la réponse en fréquence, comme exposé au paragraphe 5.16 pour un filtre à deux dimensions. Cependant, il peut être avantageux, notamment pour la précision des calculs dans le cas d'un nombre important de coefficients, de procéder dans le domaine des fréquences et en partant d'une solution approchée. De plus, cette méthode est générale et s'applique aux fonctions-coût non quadratiques, par itération; une telle approche peut s'utiliser pour le calcul des coefficients des filtres RII par exemple.

La Transformée de Fourier Discrète appliquée à la suite h_i, avec $(0 \leqslant i \leqslant N-1)$, fournit une suite H_k telle que :

$$H_k = \frac{1}{N} \sum_{i=0}^{N-1} h_i \, e^{-j2\pi \frac{ik}{N}} \tag{5.21}$$

L'ensemble des H_k, $0 \leqslant k \leqslant N-1$, constitue un échantillonnage de la réponse en fréquence du filtre avec le pas $\frac{f_e}{N}$.

Réciproquement les coefficients h_i sont liés à l'ensemble des H_k par la relation :

$$h_i = \sum_{k=0}^{N-1} H_k \, e^{j2\pi \frac{ik}{N}} \tag{5.22}$$

Par suite le problème du calcul des N coefficients est équivalent au problème de la détermination de la réponse en fréquence du filtre en N points de l'intervalle $(0, f_e)$. La fonction $H(f)$ est ensuite obtenue par la formule d'interpolation qui exprime le produit de convolution de la suite d'échantillons $H_k \, \delta\left(f - \frac{k}{N} f_e\right)$ par la transformée de Fourier de la fenêtre rectangulaire échantillonnée, calculée au paragraphe 2.4.

$$H(f) = \sum_{k=0}^{N-1} H_k \, \frac{\sin\left[\pi N\left(\frac{f}{f_e} - \frac{k}{N}\right)\right]}{N \sin\left[\pi\left(\frac{f}{f_e} - \frac{k}{N}\right)\right]} \tag{5.23}$$

On peut remarquer que cette expression constitue un autre type de développement en série de la fonction $H(f)$, à nombre limité de termes.

La fonction à approcher $D(f)$ étant donnée, une première possibilité consiste à choisir les H_k tels que :

$$H_k = D\left(\frac{k}{N} \cdot f_e\right) \quad \text{pour} \quad 0 \leqslant k \leqslant N-1$$

C'est la méthode dite de l'échantillonnage en fréquence.

La fonction de transfert du filtre $H(f)$, obtenue par interpolation, présente des ondulations en bandes passante et affaiblie, comme le montre la figure 5.14.

FIG. 5.14. *Fonction de transfert interpolée*

L'écart entre cette fonction et celle qui est donnée représente une erreur $e(f) = \mathrm{H}(f) - \mathrm{D}(f)$ qu'il est possible de minimiser au sens des moindres carrés. La procédure commence par une évaluation de l'erreur quadratique E qui est la norme L_2 de la fonction d'écart. A cet effet la réponse $\mathrm{H}(f)$ est échantillonnée avec un pas de fréquence Δ inférieur à $\frac{f_e}{\mathrm{N}}$, de façon à apparaître les valeurs interpolées, par exemple :

$$\Delta = \frac{f_e}{\mathrm{NL}} \quad \text{avec L entier supérieur à 1.}$$

La fonction $e(f)$ est calculée aux fréquences multiples de Δ.

En général dans l'évaluation de l'erreur quadratique E une partie seulement de la bande $\left(0, \frac{f_e}{2}\right)$ est à prendre en compte : pour un filtre passe-bas ce peut être la bande passante, la bande affaiblie ou l'ensemble des deux. Pour exposer le principe du calcul on suppose que la minimisation porte sur la bande passante $(0, f_1)$ d'un passe-bas; il vient dans cette hypothèse :

$$\mathrm{E} = \sum_{n=0}^{\mathrm{N}_0-1} e^2\left(n\,\frac{f_e}{\mathrm{NL}}\right) \quad \text{avec} \quad \frac{f_1}{f_e}\,\mathrm{NL} < \mathrm{N}_0 \leqslant \frac{f_1}{f_e}\,\mathrm{NL} + 1$$

De plus il est souvent utile d'affecter un coefficient de pondération $\mathrm{P}_0(n)$ à l'élément d'erreur d'indice n, afin de pouvoir modeler la réponse en fréquence; on obtient alors :

$$\mathrm{E} = \sum_{n=0}^{\mathrm{N}_0-1} \mathrm{P}_0^2(n)\, e^2\left(n\,\frac{f_e}{\mathrm{NL}}\right) = \sum_{n=0}^{\mathrm{N}_0-1} \mathrm{P}_0^2(n)\, e^2(n) \tag{5.24}$$

La fonction erreur étant obtenue à partir de la formule d'interpolation (5.23), l'erreur quadratique E est fonction de l'ensemble des H_k avec $0 \leqslant k \leqslant \mathrm{N}-1$ et est exprimée par : E(H). Si l'on donne à ces échantillons de la réponse en fréquence des accroissements $\Delta\mathrm{H}_k$, on obtient une nouvelle valeur de l'erreur quadratique qui s'exprime par l'égalité :

$$\mathrm{E}(\mathrm{H}+\Delta\mathrm{H}) = \mathrm{E}(\mathrm{H}) + \sum_{k=0}^{\mathrm{N}-1} \frac{\partial \mathrm{E}}{\partial \mathrm{H}_k}\,\Delta\mathrm{H}_k + \frac{1}{2}\sum_{k=0}^{\mathrm{N}-1}\sum_{l=0}^{\mathrm{N}-1} \frac{\partial^2 \mathrm{E}}{\partial \mathrm{H}_k\,\partial \mathrm{H}_l}\,\Delta\mathrm{H}_k\,\Delta\mathrm{H}_l \tag{5.25}$$

Compte tenu de la relation de définition de E, et de la relation d'interpolation (5.23) il vient :

$$\frac{\partial \mathrm{E}}{\partial \mathrm{H}_k} = 2\sum_{n=0}^{\mathrm{N}_0-1} \mathrm{P}_0^2(n)\, e(n)\, \frac{\partial e(n)}{\partial \mathrm{H}_k}$$

$$\frac{\partial^2 \mathrm{E}}{\partial \mathrm{H}_k\,\partial \mathrm{H}_l} = 2\sum_{n=0}^{\mathrm{N}_0-1} \mathrm{P}_0^2(n)\, \frac{\partial e(n)}{\partial \mathrm{H}_l}\cdot\frac{\partial e(n)}{\partial \mathrm{H}_k}$$

Ces équations s'écrivent sous une forme matricielle ; soit A la matrice à N lignes et N_0 colonnes telles que :

$$A = \begin{bmatrix} a_{0\,0} & a_{0\,1} & \cdots & a_{(N_0-1)\,0} \\ a_{1\,0} & a_{1\,1} & \cdots & a_{(N_0-1)\,1} \\ \vdots & \vdots & & \vdots \\ a_{0\,(N-1)} & a_{1\,(N-1)} & \cdots & a_{(N-1)\,(N_0-1)} \end{bmatrix} \quad \text{avec} \quad a_{ij} = \frac{\partial e(j)}{\partial H_i}$$

Soit P_0 la matrice diagonale d'ordre N_0 dont les éléments sont les coefficients de pondération $P_0(n)$; il vient :

$$\left[\frac{\partial E}{\partial H_k}\right] = 2\,AP_0^2\,[e(n)] \tag{5.26}$$

L'ensemble des termes $\frac{\partial^2 E}{\partial H_k\,\partial H_l}$ constitue une matrice carrée d'ordre N telle que :

$$\frac{\partial^2 E}{\partial H_k\,\partial H_l} = 2\,AP_0^2\,A^t \tag{5.27}$$

La condition pour que $E(H + \Delta H)$ soit le minimum de la fonction est que toutes ses dérivées par rapport aux $H_k\,(0 \leqslant k \leqslant N-1)$ s'annulent en ce point. Or :

$$\frac{\partial}{\partial H_k}\,E(H+\Delta H) = \frac{\partial E}{\partial H_k} + \sum_{l=0}^{N-1} \frac{\partial E}{\partial H_k}\cdot\frac{\partial E}{\partial H_l}\,\Delta H_l$$

La condition des moindres carrés s'écrit alors :

$$AP_0^2\,[e(n)] + AP_0^2\,A^t\,[\Delta H] = 0 \tag{5.28}$$

Dans ces conditions les accroissements $\Delta H_k\,(0 \leqslant k \leqslant N-1)$ qui permettent de passer des valeurs initiales des échantillons de la réponse en fréquence, aux valeurs optimales forment un vecteur colonne qui s'écrit :

$$[\Delta H] = -\,[AP_0^2\,A^t]^{-1}\,AP_0^2\,[e(n)] \tag{5.29}$$

Finalement le calcul des coefficients du filtre par l'approche proposée pour la méthode des moindres carrés demande les opérations suivantes :

1. Échantillonner la fonction à approcher en N points pour obtenir N nombres $H_k\,(0 \leqslant k \leqslant N-1)$.

2. Dans la bande de fréquences où l'erreur doit être minimisée, interpoler la réponse entre les H_k pour obtenir N_0 nombres $e(n)$ $(0 \leqslant n \leqslant N_0 - 1)$ qui représentent l'écart entre la réponse du filtre et la fonction à approcher.

3. En fonction des contraintes de l'approximation déterminer N_0 coefficients de pondération $P_0(n)$.

4. Calculer à l'aide de l'équation d'interpolation les éléments de la matrice A.
5. Résoudre l'équation matricielle qui donne les ΔH_k.

6. Opérer sur l'ensemble des nombres $(H_k + \Delta H_k)$ avec $0 \leqslant k \leqslant N - 1$ une transformation de Fourier inverse pour obtenir les coefficients du filtre.

Les coefficients de pondération $P_0(n)$ permettent, par exemple, d'obtenir des ondulations en bandes passantes et affaiblies qui soient dans un rapport donné ou encore d'imposer à la réponse en fréquence de passer par un point particulier ; cette dernière condition peut aussi être prise en compte par la réduction d'une unité du nombre de degrés de liberté, ce qui est plus élégant mais plus compliqué à programmer.

La mise en œuvre de la procédure de calcul ne présente pas de difficultés particulières ; elle permet de calculer un filtre d'une manière directe. Cependant le filtre obtenu a des ondulations qui n'ont pas une amplitude constante ; or c'est un objectif qui se rencontre fréquemment. Pour l'atteindre il faut faire appel à une technique itérative.

Si l'inversion de matrice de la relation 5.29 est délicate ou impossible, il est possible d'atteindre l'optimum en remplaçant cette matrice par une constante faible et en itérant le processus, c'est l'algorithme du gradient.

5.5 CALCUL DES COEFFICIENTS PAR TFD

Une première approche itérative consiste à utiliser la Transformation de Fourier Discrète, qui se calcule efficacement par un algorithme rapide.

Soit à calculer un filtre à phase linéaire à N coefficients et satisfaisant au gabarit de la figure 5.7. On va utiliser une transformée de Fourier Discrète d'ordre N_0 avec $N_0 \simeq 10\,N$.

La procédure consiste à prendre des valeurs initiales pour les coefficients, par exemple les termes h donnés par (5.18), pour $-P \leqslant i \leqslant P$, si $N = 2P + 1$. Cet ensemble de N valeurs est complété symétriquement par des zéros pour obtenir un ensemble de N_0 valeurs réelles, symétriques par rapport à l'origine.

Ensuite, un calcul de TFD donne la réponse $H(f)$ en N_0 points de l'axe des fréquences. On peut écrire :

$$H(f) = H_{id}(f) + E(f)$$

où $H_{id}(f)$ est la réponse idéale et $E(f)$ l'écart par rapport à cette réponse. On effectue alors un écrêtage de l'écart $E(f)$, c'est-à-dire que l'on remplace $H(f)$ par la fonction $G(f)$ telle que :

$$G(f) = H_{id}(f) + E_L(f) \quad \text{si} \quad H(f) > H_{id}(f) + E_L(f)$$
$$G(f) = H_{id}(f) - E_L(f) \quad \text{si} \quad H(f) < H_{id}(f) - E_L(f)$$

où $E_L(f)$ représente la limite de l'écart donnée par le gabarit, par exemple δ_1 ou δ_2 pour le filtre passe-bas de la figure 5.7.

Un calcul de TFD inverse donne N_0 termes dont on conserve les N valeurs qui encadrent l'origine, en annulant les autres. Puis, la procédure recommence, en prenant la TFD des N_0 valeurs ainsi obtenues.

En désignant par $J(k)$ la somme des carrés des $N_0 - N$ termes annulés dans le domaine temporel à l'itération k, on obtient une fonction décroissante si les spécifications du filtre sont compatibles avec le nombre N de coefficients. On arrête la procédure quand $J(k)$ tombe au-dessous d'un seuil fixé.

En appliquant la méthode pour différents nombres de coefficients N, on peut approcher la solution optimale et même l'atteindre dans des cas particuliers. Tous les types de filtres à phase linéaire peuvent se calculer ainsi.

Pour obtenir le filtre optimal, une méthode basée sur l'approximation de Tchebycheff est utilisée.

5.6 CALCUL DES COEFFICIENTS PAR APPROXIMATION DE TCHEBYCHEFF

Le but à atteindre est d'obtenir un filtre dont la réponse en fréquence présente des ondulations d'amplitude constante, de manière à approcher au mieux un gabarit, comme celui qui est donné sur la figure 5.7 pour un passe-bas dont les ondulations ne doivent pas dépasser l'amplitude δ_1 en bande passante et δ_2 en bande affaiblie. C'est un problème qui relève de l'approximation d'une fonction par un polynôme au sens de Tchebycheff, la norme à considérer pour la fonction d'écart est la norme L_∞.

D'après l'expression de la fonction de transfert d'un filtre RIF à phase linéaire, le calcul des coefficients se ramène à la détermination de la fonction $H_R(f)$ qui s'écrit :

$$H_R(f) = \sum_{i=0}^{r-1} h_i \cos(2\pi f i T) \tag{5.30}$$

quand le nombre de coefficients s'élève à : $N = 2r - 1$. La technique qui va être présentée est valable dans tous les cas, que N soit pair ou impair, que les coefficients soient symétriques ou antisymétriques. Elle est basée sur le théorème d'analyse numérique suivant [1] :

Théorème : Une condition nécessaire et suffisante pour que $H_R(f)$ soit l'unique et meilleure approximation au sens de Tchebycheff d'une fonction donnée $D(f)$ sur un sous-ensemble compact A de l'intervalle [0, 1/2], est que la fonction erreur $e(f) = H_R(f) - D(f)$ présente au moins $(r + 1)$ fréquences extrémales sur A $(f_0, f_1, \ldots, f_r)$, telles que $e(f_i) = -e(f_{i-1})$ avec $1 \leqslant i \leqslant r$ et

$$|e(f_i)| = \max_{f \in A} |e(f)|$$

Ce résultat reste valable si une fonction de pondération $P_0(f)$ de l'erreur est introduite.

Le problème se trouve ainsi ramené à la résolution du système de $(r + 1)$ équations :

$$P_0(f_i)\,[D(f_i) - H_R(f_i)] = (-1)^i\,\delta$$

Les inconnues sont les coefficients du filtre $h_i (0 \leqslant i \leqslant r - 1)$ et le maximum de la fonction erreur : δ.

Sous forme matricielle en faisant apparaître les inconnues dans un vecteur colonne et en normalisant les fréquences de manière que $f_e = \frac{1}{T} = 1$ il vient :

$$\begin{bmatrix} D(f_0) \\ D(f_1) \\ \vdots \\ D(f_r) \end{bmatrix} = \begin{bmatrix} 1 & \cos(2\pi f_0)\ldots & \cos[2\pi f_0(r-1)] & \dfrac{1}{P_0(f_0)} \\ 1 & \cos(2\pi f_1)\ldots & \cos[2\pi f_1(r-1)] & \dfrac{-1}{P_0(f_1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \cos(2\pi f_r)\ldots & \cos[2\pi f_r(r-1)] & \dfrac{(-1)^r}{P_0(f_r)} \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_{r-1} \\ \delta \end{bmatrix}$$

Cette équation matricielle conduit à la détermination des coefficients du filtre, à la condition cependant que soient connues les $(r + 1)$ fréquences extrémales f_i.

C'est dans la recherche des fréquences extrémales qu'intervient une procédure itérative réalisée suivant un algorithme dit de Remez et dont chaque étape comprend les phases suivantes :

– Des valeurs initiales sont affectées, ou sont disponibles pour les paramètres $f_i (0 \leqslant i \leqslant r)$.

– La valeur δ correspondante est calculée en résolvant le système d'équations, ce qui conduit à la formule suivante :

$$\delta = \frac{a_0 D(f_0) + a_1 D(f_1) + \ldots + a_r D(f_r)}{a_0/P_0(f_0) - a_1/P_0(f_1) + \ldots + (-1)^r\, a_r/P_0(f_r)}$$

avec :

$$a_k = \prod_{\substack{i=0 \\ i \neq k}}^{r} \frac{1}{\cos(2\pi f_k) - \cos(2\pi f_i)}$$

– Les valeurs de la fonction $H_R(f)$ sont interpolées entre les $f_i (0 \leqslant i \leqslant r)$ pour calculer $e(f)$.

– Les fréquences extrémales obtenues sont prises comme valeurs initiales pour l'étape suivante.

La figure 5.15 montre l'évolution de la fonction erreur dans une étape du calcul. La procédure est arrêtée quand la différence entre la valeur δ calculée au moyen

des nouvelles fréquences extrémales et la valeur précédente tombe au-dessous d'un seuil fixé à l'avance. Ce résultat est obtenu en quelques itérations dans la grande majorité des cas.

La convergence de cette procédure est liée au choix des valeurs initiales des fréquences f_i; pour la première itération, on peut prendre les fréquences extrémales obtenues avec une autre méthode de calcul pour les coefficients du filtre, ou même simplement, une répartition uniforme des fréquences extrémales sur l'intervalle de fréquence considéré.

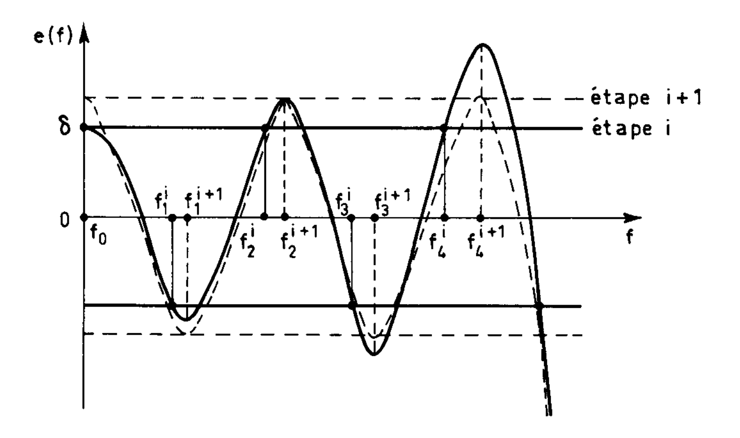


FIG. 5.15. *Évolution de la fonction erreur dans une étape de l'algorithme de Remez*

Comme dans la méthode des moindres carrés du paragraphe précédent, on trouve une étape d'interpolation des valeurs de $H_R(f)$ qui, du fait de la répartition non uniforme des fréquences extrémales, est plus commode à réaliser en faisant appel aux formules d'interpolation de Lagrange :

$$H_R(f) = \frac{\sum_{k=0}^{r-1} \frac{\beta_k}{(x - x_k)} \left[D(f_k) - (-1)^k \frac{\delta}{P_0(f_k)} \right]}{\sum_{k=0}^{r-1} \frac{\beta_k}{(x - x_k)}} \tag{5.31}$$

avec :

$$\beta_k = \prod_{\substack{i=0 \\ i \neq k}}^{r-1} \frac{1}{x_k - x_i} \quad \text{et} \quad x = \cos(2\pi f)$$

A la fin de la procédure d'itération, les fréquences extrémales obtenues sont utilisées pour produire un échantillonnage à pas constant de la réponse en fré-

quence, qui, par transformation de Fourier discrète inverse, fournit les coefficients du filtre.

Des filtres ayant plusieurs centaines de coefficients peuvent être calculés suivant cette technique, qui s'applique aux filtres passe-bas, passe-haut et passe-bande, avec ou sans déphasage fixe [2]. Un exemple de calcul est donné en annexe.

5.7 RELATIONS ENTRE NOMBRE DE COEFFICIENTS ET GABARIT DE FILTRE

Dans les techniques de calcul qui ont été exposées le nombre de coefficients N du filtre a été supposé donné à priori. Or dans la pratique N est un paramètre important, par exemple dans les projets où il faut évaluer la capacité de calcul nécessaire à la mise en œuvre d'un filtre numérique satisfaisant à un gabarit donné.

Pour un filtre passe-bas, comme indiqué sur la figure 5.7, le gabarit est donné par l'ondulation en bandes passantes et affaiblie δ_1 et δ_2, la fréquence marquant la fin de la bande passante f_1, et la bande de transition $\Delta f = f_2 - f_1$. En analysant les résultats du calcul d'un grand nombre de filtres avec des spécifications très variées, on constate qu'en première approximation le nombre de coefficients est proportionnel au logarithme de $\frac{1}{\delta_1}$ et de $\frac{1}{\delta_2}$ ainsi qu'au rapport de la fréquence d'échantillonnage fe à la bande de transition Δf. Par ajustage des paramètres on obtient alors l'estimation Ne suivante pour le nombre de coefficients :

$$Ne = \frac{2}{3} \log\left[\frac{1}{10 \,.\, \delta_1 \,.\, \delta_2}\right] . \frac{fe}{\Delta f} \qquad (5.32)$$

Cette estimation particulièrement simple est suffisante dans la plupart des cas rencontrés en pratique. Elle met bien en évidence l'importance relative des paramètres. La bande de transition Δf est le paramètre le plus sensible ; les ondulations en bandes passante et affaiblie ont une contribution secondaire : par exemple quand $\delta_1 = \delta_2 = 0{,}01$, une division par deux de l'une de ces valeurs entraîne seulement une augmentation de 10 % de l'ordre du filtre. De plus, il est remarquable de constater que, selon cette évaluation, la complexité du filtre ne dépend pas de la largeur de la bande passante.

Exemples

1. Le calcul d'un filtre à 39 coefficients (N = 39) a conduit aux valeurs suivantes :

$$\delta_1 = 0{,}017 ; \quad \delta_2 = 0{,}034 ; \quad f_1 = 0{,}10375 ;$$
$$f_2 = 0{,}14375 ; \quad \Delta f = 0{,}04$$

Avec ces valeurs de paramètres l'estimation donne : Ne = 40.

2. Un filtre à 160 coefficients (N = 160) a comme paramètres :

$$\delta_1 = 2{,}24.10^{-2};\quad \delta_2 = 1{,}12.10^{-4};\quad f_1 = 0{,}053125;$$
$$f_2 = 0{,}071875;\quad \Delta f = 0{,}01875$$

L'estimation donne N*e* = 164.

3. Un filtre à 15 coefficients (N = 15) a comme paramètres :

$$\delta_1 = 0{,}0411;\quad \delta_2 = 0{,}0137;\quad f_1 = 0{,}1725;$$
$$f_2 = 0{,}2875;\quad \Delta f = 0{,}115$$

L'estimation donne : N*e* = 13.

Les coefficients du filtre correspondant à l'équation :

$$y(n) = \sum_{i=1}^{15} a_i x(n-i)$$

ont pour valeur :

$$\begin{aligned}
a_1 &= -\,0{,}00047 = a_{15}\\
a_2 &= 0{,}02799 = a_{14}\\
a_3 &= 0{,}02812 = a_{13}\\
a_4 &= -\,0{,}03572 = a_{12}\\
a_5 &= -\,0{,}07927 = a_{11}\\
a_6 &= 0{,}04720 = a_{10}\\
a_7 &= 0{,}30848 = a_9\\
a_8 &= 0{,}44847
\end{aligned}$$

Les ondulations du filtre sont données sur la figure 5.16.

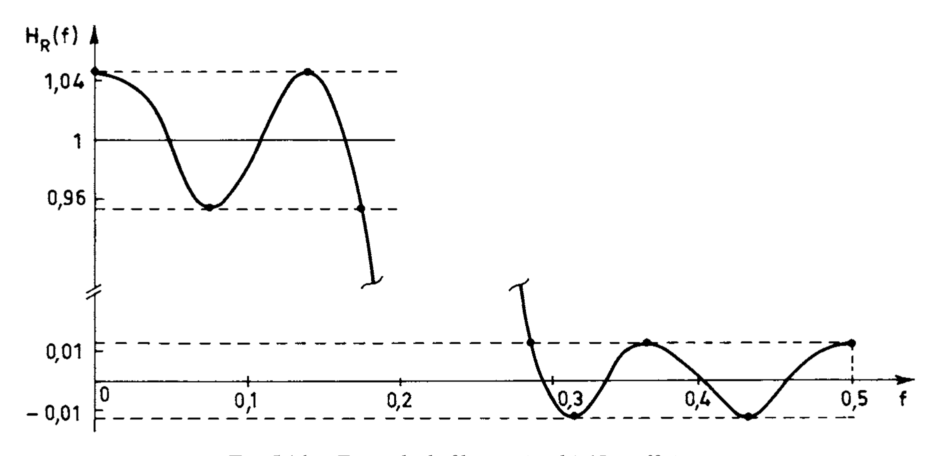


FIG. 5.16. *Exemple de filtre optimal à 15 coefficients*

Il faut cependant noter que des écarts non négligeables peuvent apparaître entre la valeur N réellement nécessaire et la valeur estimée N*e*, quand les limites

de la bande de transition approchent les valeurs 0 et 0,5 ou encore quand N prend des valeurs de quelques unités. Un ensemble de formules plus élaborées est donné dans la référence [3].

Comme indiqué aux chapitres 7 et 10 un filtre passe-haut peut être obtenu à partir d'un passe-bas en inversant le signe d'un coefficient sur deux. Il s'en suit que l'estimation (5.32) s'applique aussi aux filtres passe-haut. Quand le gabarit présente pour les bandes passante et affaiblie des plages de fréquences où les ondulations doivent être différentes, un majorant du nombre de coefficients peut être obtenu en prenant pour δ_1 et δ_2 les contraintes les plus sévères en bandes passantes et affaiblies respectivement.

Dans le cas des filtres passe-bande, il faut faire intervenir plusieurs bandes de transition. La figure 5.17 donne le gabarit d'un tel filtre ayant deux bandes de transition Δf_1 et Δf_2. L'expérience montre que le nombre de coefficients N dépend essentiellement de la bande de transition la plus faible. $\Delta f_m = \min(\Delta f_1, \Delta f_2)$. On peut alors appliquer l'estimation (5.32) avec $\Delta f = \Delta f_m$. Un majorant pour le nombre de coefficients est obtenu en considérant le filtre passe-bande comme la mise en cascade d'un filtre passe-bas et d'un passe-haut et en faisant la somme des estimations.

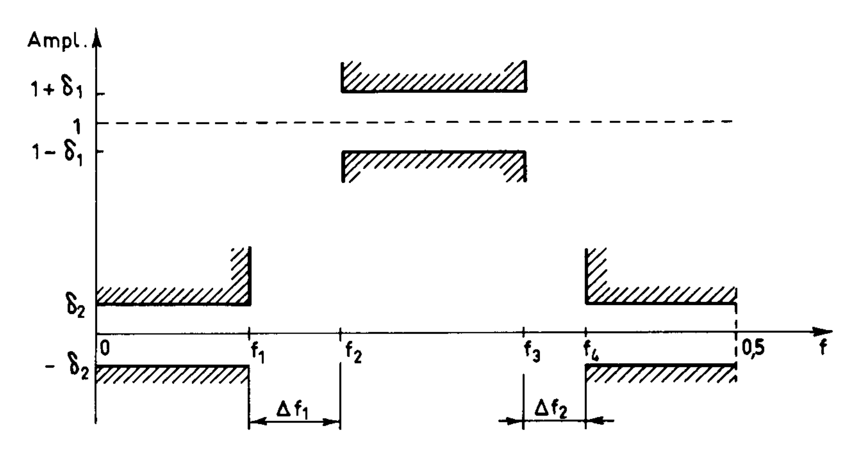


FIG. 5.17. *Gabarit d'un filtre passe-bande*

Exemple

Un filtre passe-bande à 32 coefficients (N = 32) présente les caractéristiques suivantes :

$$\delta_1 = 0{,}015\,; \quad \delta_2 = 0{,}0015\,; \quad f_1 = 0{,}1\,; \quad f_2 = 0{,}2\,;$$
$$f_3 = 0{,}35\,; \quad f_4 = 0{,}425\,; \quad \Delta f_m = 0{,}075.$$

L'estimation par la relation (5.32) avec $\Delta f = \Delta f_m$ donne Ne = 32.

Un ensemble de formules pour l'estimation de l'ordre des filtres passe-bande est donné dans la référence [3].

Les formules d'estimation peuvent être utilisées pour compléter le programme de calcul des coefficients du filtre, en faisant déterminer le nombre N en début de programme. La relation (5.32) est très utile dans les projets, pour les évaluations de complexité.

Quand on observe les réponses en fréquence des filtres calculés dans la bande de transition, on remarque qu'elles sont proches d'une cosinusoïde surélevée et d'autant plus que les ondulations en bandes passante et affaiblie sont proches. En fait, ce type de réponse correspond aux spécifications imposées en transmission de données avec les filtres dits de Nyquist et il représente une autre approche des filtres RIF à phase linéaire.

5.8 FILTRE À TRANSITION EN COSINUS SURÉLEVÉ ET COSINUS FILTRE DE NYQUIST – FILTRE DEMI-BANDE

La réponse en fréquence H(f) d'un filtre dont la bande de transition est en cosinus surélevé est représentée à la figure 5.18.*a*.

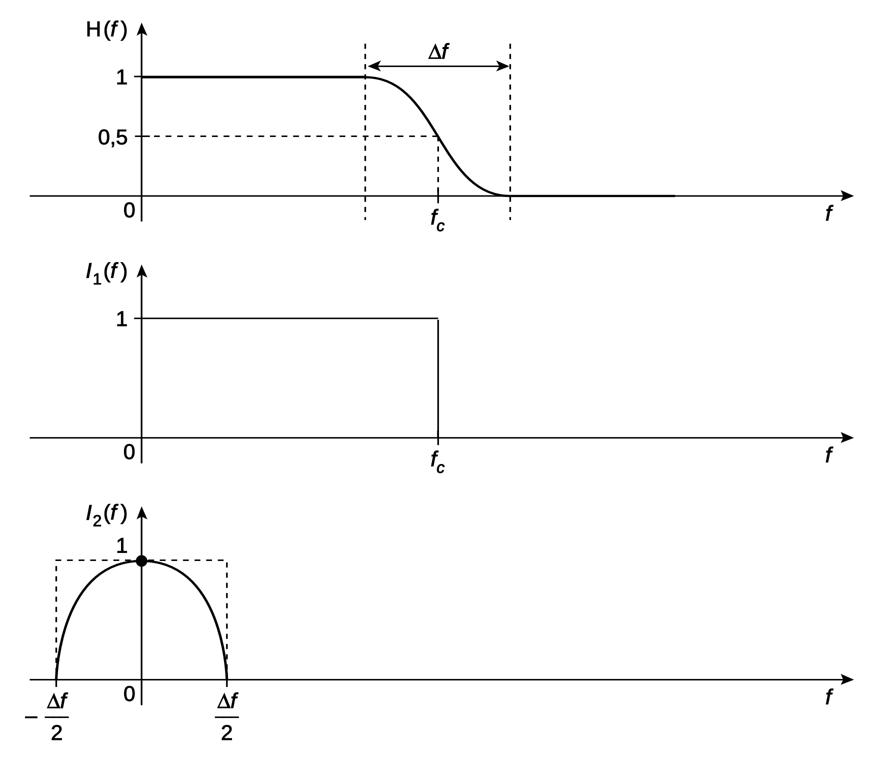


FIG. 5.18.

a) *réponse avec transition en cosinus surélevé,*
b) *impulsion en fréquence de largeur* $2f_c$,
c) *impulsion en fréquence pour la transition.*

On vérifie qu'elle s'exprime comme le produit de convolution suivant :

$$\mathrm{H}(f) = \mathrm{I}_1(f) * \left[\mathrm{I}_2(f) \,.\, \frac{\pi}{2\Delta f} \cos\left(\frac{\pi f}{\Delta f}\right)\right] \qquad (5.33)$$

où $\mathrm{I}_1(f)$ est une impulsion de largeur $2f_c$ et $\mathrm{I}_2(f)$ une impulsion de largeur Δf.

Dans ces conditions, la réponse impulsionnelle $h(t)$ s'écrit comme le produit de 2 réponses impulsionnelles $i_1(t)$ et $i_2(t)$ données par :

$$i_1(t) = 2\,f_c\ \frac{\sin \pi\, 2 f_c t}{\pi\, 2 f_c t}$$

et

$$i_2(t) = \frac{1}{2}\left(\frac{\sin \pi \Delta f\left(t + \dfrac{1}{2\Delta f}\right)}{2\Delta f\left(t + \dfrac{1}{2\Delta f}\right)} + \frac{\sin \pi \Delta f\left(t - \dfrac{1}{2\Delta f}\right)}{2\Delta f\left(t - \dfrac{1}{2\Delta f}\right)}\right)$$

Après simplifications, il vient :

$$h(t) = 2 f_c\ \frac{\sin 2\pi f_c t}{2\pi f_c t} \cdot \frac{\cos \pi\, \Delta f t}{1 - 4\Delta f^2 t^2} \qquad (5.34)$$

Le nombre total de coefficients du filtre est déterminé principalement par la fonction $i_2(t)$ et la largeur de son lobe principal, égale à $3/\Delta f$.

Ainsi dans un filtre à bande de transition en cosinus surélevé, le nombre de coefficients peut être estimé par :

$$\mathrm{N} \simeq \frac{3fe}{\Delta f} \qquad (5.35)$$

Cette estimation peut être considérée comme une première approche, quand on la compare à la relation (5.32).

Les coefficients du filtre numérique sont obtenus par échantillonnage de $h(t)$, pour $|i| \leqslant \mathrm{P}$ et $\mathrm{N} = 2\mathrm{P} + 1$, soit :

$$h_i = \frac{2 f_c}{f_e}\ \frac{\sin 2\pi \dfrac{f_c}{f_e}\, i}{2\pi \dfrac{f_c}{f_e}\, i}\ \frac{\cos \pi \dfrac{\Delta f}{f_e}\, i}{1 - 4(\Delta f / f_e)^2 i^2} \qquad (5.36)$$

À noter que cette expression peut être appliquée à un filtre quelconque et donne une estimation directe des coefficients plus précise que la relation (5.18).

Les résultats ci-dessus se généralisent à toute bande de transition possédant la propriété de symétrie, c'est-à-dire que :

$$H(f_c + f) = 1 - H(f_c - f)\,;\ |f| \leqslant \frac{\Delta f}{2}$$

Ces filtres sont à la base de la transmission numérique et on les désigne par filtres « de Nyquist ». Leur réponse impulsionnelle s'annule à tous les instants multiples de $1/2f_c$ et, en traitement numérique, les coefficients d'indice multiple de $f_e/2f_c$ sont nuls. Un cas particulier important est celui du filtre demi-bande, dans lequel $f_c = f_e/4$. Alors, les coefficients pairs s'annulent et, pour $N = 4M + 1$ coefficients, la relation d'entrée-sortie s'écrit :

$$y(n) = \frac{1}{2}\left[x(n-2M) + \sum_{i=1}^{M} h_{2i-1}[x(n-2M+2i-1) + x(n-2M-2i+1]\right]$$

Pour la réponse en fréquence, il vient :

$$H(f) = e^{-j2\pi 2Mf}\,\frac{1}{2}\left[1 + 2\sum_{i=1}^{M} h_{2i-1}\cos 2\pi(2i-1)f\right]$$

Ce filtre nécessite une quantité de calculs réduite et c'est un élément de base du filtrage multicadence.

En transmission, la fonction de filtrage se partage entre l'émetteur et le récepteur et on utilise fréquemment le filtre demi-Nyquist, par exemple avec une bande de transition en cosinus :

$$H^{1/2}(f) = 1\,; \quad |f| \leqslant f_c - \frac{\Delta f}{2}$$

$$H^{1/2}(f) = \cos \pi\left[f - \left(f_c - \frac{\Delta f}{2}\right)\right]/2\Delta f\,; \quad f_c - \frac{\Delta f}{2} \leqslant f \leqslant f_c + \frac{\Delta f}{2}$$

$$H^{1/2}(f) = 0\,; \quad |f| \geqslant f_c + \frac{\Delta f}{2}$$

La réponse impulsionnelle s'écrit :

$$h^{1/2}(t) = \frac{\dfrac{4\Delta f}{\pi}\cos 2\pi t\left(f_c + \dfrac{\Delta f}{2}\right) + \dfrac{1}{\pi t}\sin 2\pi t\left(f_c - \dfrac{\Delta f}{2}\right)}{1 - (4\Delta f t)^2} \tag{5.37}$$

Comme précédemment, un filtre numérique « demi-Nyquist » peut être obtenu en échantillonnant cette fonction, c'est-à-dire en remplaçant t par i/f_e dans la relation (5.37).

5.9 STRUCTURES POUR LA RÉALISATION DES FILTRES RIF

La mise en œuvre des filtres RIF se fait par des circuits qui réalisent les trois opérations fondamentales que sont la mise en mémoire, la multiplication et l'addition et qui sont agencés pour fournir, à partir de la suite des données $x(n)$, une suite de

sortie $y(n)$ conformément à l'équation de définition du filtre. Aucune opération réelle n'étant instantanée, l'équation qui est réalisée à la place de la relation (5.5) du paragraphe 5.1 est la suivante :

$$y(n) = \sum_{i=0}^{N-1} a_i x(n-i-1) \tag{5.38}$$

Il faut N mémoires de données et pour chaque nombre de sortie il faut faire N multiplications et N – 1 additions. Différents arrangements de circuits peuvent être envisagés pour mettre en œuvre ces opérations [4, 5, 6].

La figure 5.19.*a* donne le schéma du filtre dans la structure dite directe. La transposition du graphe de ce schéma conduit à la structure dite transposée et représentée sur la figure 5.19.*b* où les mêmes opérateurs sont agencés différemment. Cette structure amène à réaliser la multiplication de chacune des données $x(n)$ par tous les coefficients successivement; d'autre part les mémoires stockent des sommes partielles; en effet au temps n la première mémoire stocke le nombre $a_{N-1}x(n)$, la suivante : $a_{N-1}x(n-1) + a_{N-2}x(n)$ et la dernière stocke la somme $y(n)$. La différence entre ces deux structures tient à la position des mémoires. On peut aussi envisager une structure intermédiaire à deux mémoires par coefficient, où les données internes sont stockées pendant la durée T/2 dans chacune; la structure en chaîne ainsi obtenue ne présente que des interconnexions locales.

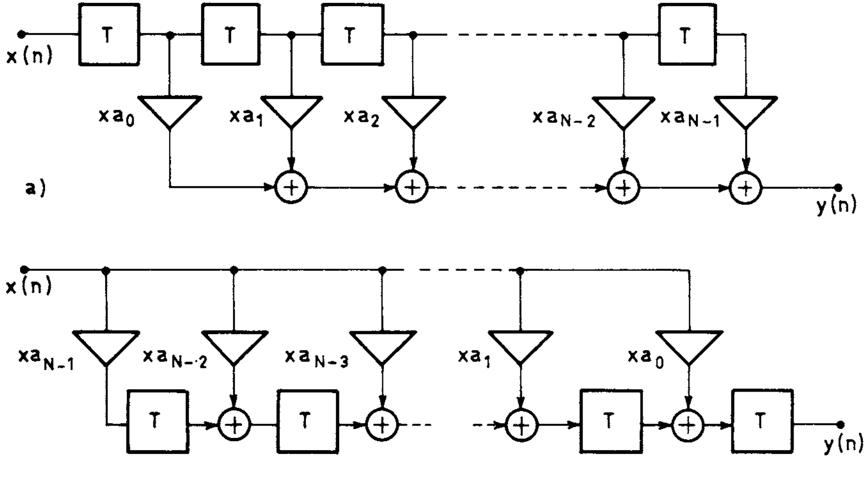


FIG. 5.19. *Réalisation des filtres RIF*
a) *structure directe,*
b) *structure transposée.*

Dans les filtres à phase linéaire, la symétrie des coefficients peut être exploitée pour diviser par deux le nombre de multiplications à faire par nombre de sortie, ce qui est très important pour la complexité du filtre et justifie l'utilisation quasi générale de filtres à phase linéaire. La structure correspondante est présentée sur la figure 5.20 pour la forme directe quand le nombre de coefficients est impair : N = 2P + 1.

La complexité des circuits dépend du nombre d'opérations à faire, mais aussi de l'ampleur de ces opérations; c'est ainsi que les termes de la multiplication doivent avoir un nombre de bits aussi réduit que possible ce qui tend à diminuer la capacité de mémoire nécessaire, tant pour les coefficients que pour les données. Ces limitations modifient les caractéristiques de traitement.

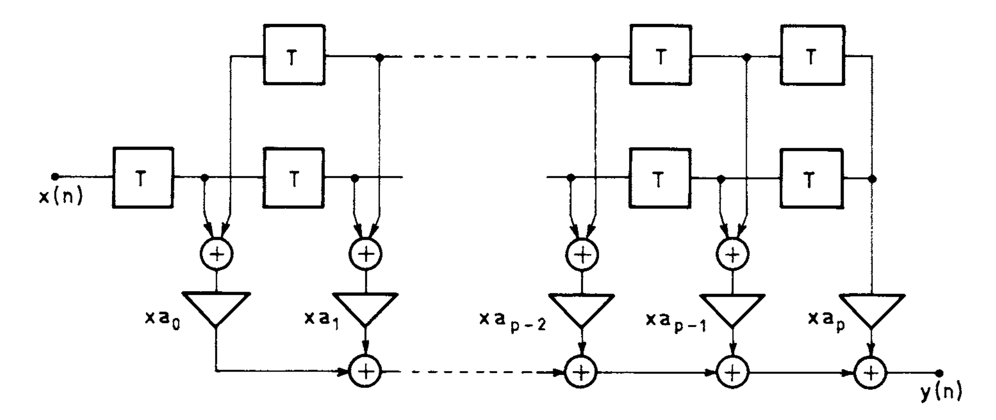


Fig. 5.20. *Structure directe pour filtre à phase linéaire*

5.10 LIMITATIONS DU NOMBRE DE BITS DES COEFFICIENTS

La limitation du nombre de bits des coefficients d'un filtre entraîne une altération de la réponse en fréquence qui se traduit par la superposition d'une fonction parasite. Les conséquences vont être analysées dans le cas des filtres à phase linéaire. L'extension des résultats aux filtres RIF quelconques ne présente pas de difficultés.

Soit un filtre à phase linéaire à $N = 2P + 1$ coefficients, dont la fonction de transfert s'écrit d'après le paragraphe 5.2, relation (5.13) :

$$H(f) = e^{-j2\pi f PT} \left[h_0 + 2 \sum_{i=1}^{P} h_i \cos(2\pi f i T) \right]$$

La limitation du nombre de bits des nombres qui représentent les coefficients se traduit par une erreur δh_i $(0 \leqslant i \leqslant P)$ sur le coefficient h_i qui, dans l'hypothèse d'un arrondi avec un échelon de quantification q, est telle que :

$$|\delta h_i| \leqslant \frac{q}{2}.$$

Il en résulte la superposition à la fonction $H(f)$ d'une fonction parasite $e(f)$ telle que :

$$e(f) = e^{-j2\pi f PT}\left[\delta h_0 + 2\sum_{i=1}^{P} \delta h_i \cos(2\pi f i T)\right] \tag{5.39}$$

L'amplitude de cette fonction doit être limitée, pour que la réponse du filtre réel reste dans le gabarit imposé.

Une borne est obtenue comme suit :

$$|e(f)| \leqslant |\delta h_0| + 2\sum_{i=1}^{P} |\delta h_i|\,|\cos(2\pi f i T)|$$

$$|e(f)| \leqslant \frac{q}{2}\,.\,N \tag{5.40}$$

Cette borne est en général beaucoup trop grande ; pour avoir une estimation plus réaliste il faut faire appel à une estimation statistique [7].

Quand l'étude porte sur un grand nombre de filtres ayant des spécifications variées et que l'on recherche des résultats généraux, on peut considérer les variables $\delta h_i (0 \leqslant i \leqslant P)$ comme aléatoires, indépendantes et à répartition uniforme sur l'intervalle $\left[-\frac{q}{2}, \frac{q}{2}\right]$; dans ces conditions elles ont comme variance : $\frac{q^2}{12}$.
La fonction $e(f)$ peut être considérée comme aléatoire également. Soit e_0 sa valeur efficace sur l'intervalle de fréquence $[0, f_e]$, c'est-à-dire telle que :

$$e_0^2 = \frac{1}{f_e}\int_0^{f_e} |e(f)|^2\, df \tag{5.41}$$

En fait la fonction $e(f)$ est une fonction périodique définie par son développement en série de Fourier et l'égalité de Bessel-Parseval permet d'écrire, conformément à la relation (5.8) :

$$\frac{1}{f_e}\int_0^{f_e} |e(f)|^2\, df = \sum_{i=0}^{N-1} (\delta h_i)^2$$

Par suite la variance σ^2 de la variable aléatoire e_0 s'écrit :

$$\sigma^2 = E[e_0^2] = N\,.\,\frac{q^2}{12}$$

En faisant l'hypothèse d'indépendance de la fréquence pour le moment du second ordre et compte tenu de la relation (5.41), la variable $e(f)$ peut être considérée comme une variable aléatoire de variance σ^2 telle que :

$$\sigma = \frac{q}{2}\sqrt{\frac{N}{3}} \tag{5.42}$$

Cette relation fournit une estimation de $|e(f)|$ beaucoup plus faible que la borne (5.40). En fait $e(f)$, résultant d'après (5.39) d'une somme pondérée de

variables supposées indépendantes, peut être assimilée à une variable gaussienne, de moyenne nulle si la quantification est faite par arrondi, et d'écart type σ. Pour déterminer l'échelon de quantification q, on peut alors raisonner avec les intervalles de confiance ; par exemple la probabilité pour que $|e(f)|$ dépasse la valeur 2σ est inférieure à 5 % d'après le tableau donné en annexe 2 au chapitre 1.

Les résultats ci-dessus vont maintenant être utilisés pour estimer le nombre de bits b_c nécessaire dans la représentation des coefficients d'un filtre spécifié par un gabarit.

Étant donné un gabarit de filtre, soit δ_m la valeur imposée par le gabarit pour l'amplitude des ondulations ; soit δ_0 l'amplitude des ondulations du filtre avant limitation du nombre de bits des coefficients. La fonction parasite $e(f)$ doit être telle que :

$$|e(f)| < \delta_m - \delta_0$$

Le degré de confiance dans l'estimation est supérieur à 95 % si q est choisi tel que :

$$\frac{q}{2}\sqrt{\frac{N}{3}} < \frac{\delta_m - \delta_0}{2}$$

Dans ces conditions :

$$q < (\delta_m - \delta_0)\sqrt{\frac{3}{N}} \qquad (5.43)$$

Le nombre de bits b_c nécessaire pour représenter les coefficients dépend de la plus grande des valeurs $h_i (0 \leqslant i \leqslant P)$ et, compte tenu du signe, l'échelon de quantification q est donné par :

$$q = 2^{1-b_c} \cdot \left[\max_{0 \leqslant i \leqslant P} |h_i|\right] \qquad (5.44)$$

Si le filtre est un passe-bas dont la réponse en fréquence approche l'unité en bande passante et correspondant au gabarit de la figure 5.7, les valeurs des coefficients peuvent en première approximation être calculées par la relation (5.18). Dans ces conditions le maximum est obtenu pour h_0 avec :

$$h_0 = \frac{f_1 + f_2}{f_e} \qquad (5.45)$$

Alors (5.43) et (5.44) conduisent à l'estimation suivante :

$$b_c \simeq 1 + \log 2\left[\frac{f_1 + f_2}{f_e} \cdot \sqrt{\frac{N}{3}} \cdot \frac{1}{\delta_m - \delta_0}\right] \qquad (5.46)$$

avec :

b_c : nombre de bits des coefficients (signe compris).
N : nombre de coefficients du filtre.
f_1 : limite de bande passante.
f_2 : début de bande affaiblie.
f_e : fréquence d'échantillonnage.
δ_m : limite imposée par le gabarit pour l'amplitude des ondulations.
δ_0 : amplitude des ondulations du filtre avant limitation du nombre de bits des coefficients.

Exemple

Soit le filtre passe-bas à 15 coefficients du paragraphe 5.7 dont les paramètres sont les suivants :

$$N = 15 \quad f_e = 1 \quad f_1 = 0{,}1725 \quad f_2 = 0{,}2875$$

Le gabarit impose :

$$\delta_1 = 0{,}05 \quad \delta_2 = 0{,}02$$

Les ondulations du filtre en bandes passante et affaiblie, avant limitation du nombre de bits des coefficients, ont pour valeurs :

$$\delta_{10} = 0{,}0411 \quad \delta_{20} = 0{,}0137$$

Dans ces conditions :

$$\delta_m - \delta_0 = \min\,(\delta_1 - \delta_{10}, \delta_2 - \delta_{20}) = 0{,}0063$$

Il vient :

$$b_c \simeq 1 + \log 2\left(0{,}46\,\frac{\sqrt{5}}{0{,}0063}\right) = 8{,}3$$

On choisit $b_c = 8$, c'est-à-dire des coefficients à 8 bits. La fonction $e(f)$ correspondante est représentée sur la figure 5.21.

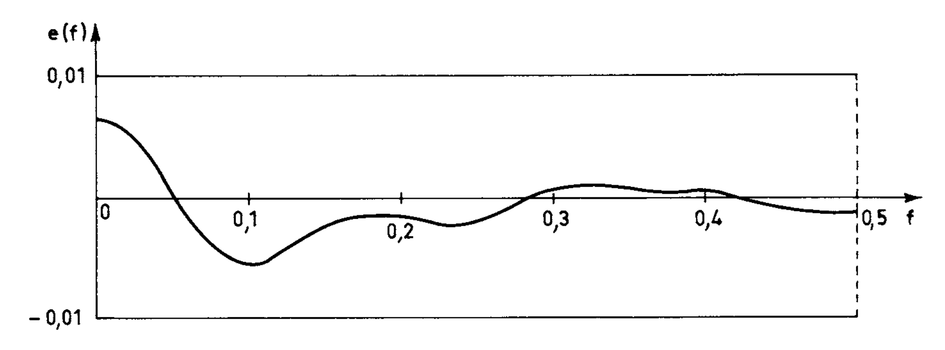


FIG. 5.21. *Erreur due à l'arrondi des coefficients.*

Dans la pratique la relation (5.46) peut être simplifiée. D'abord la tolérance fournie par le gabarit du filtre est généralement répartie équitablement entre les ondulations avant limitation du nombre de bits des coefficients et l'erreur supplémentaire due à cette limitation, c'est-à-dire que $\delta_0 = \delta_m/2$. De plus, les filtres à réaliser sont généralement tels que :

$$\frac{f_e}{\Delta f}\,.\,0{,}5 \leqslant \frac{N}{3} \leqslant \frac{f_e}{\Delta f}\,.\,1{,}5 \qquad (5.47)$$

ce qui, d'après la relation (5.32) correspond à une gamme importante des valeurs des paramètres δ_1 et δ_2. Dans ces conditions on peut se baser sur l'estimation (5.35) et remplacer $\frac{N}{3}$ par $\frac{f_e}{\Delta f}$ et une estimation convenable du nombre de bits des coefficients est donnée par :

$$b_c \simeq 1 + \log 2 \left[\frac{f_1 + f_2}{f_e} \cdot \sqrt{\frac{f_e}{\Delta f}} \cdot \frac{2}{\delta_m} \right].$$

En faisant apparaître la raideur de coupure du filtre (5.17) et la bande de transition normalisée, on obtient finalement :

$$b_c \simeq 3 + \log 2 \left(\frac{f_1 + f_2}{2\Delta f} - \frac{1}{2} \log 2 \left(\frac{f_e}{\Delta f} \right) + \log 2 \left(\frac{1}{\min\{\delta_1, \delta_2\}} \right) \qquad (5.48)$$

Ainsi le nombre de bits des coefficients est directement lié aux spécifications du filtre. Il est remarquable de constater que les filtres à bande passante étroite demandent moins de bits que les filtres à bande large.

Les relations (5.46) et (5.48) s'appliquent aux filtres passe-haut et elles s'étendent aux filtres passe-bande.

L'analyse ci-dessus a été menée avec l'hypothèse d'une séparation des opérations de calcul des coefficients et de limitation de leur nombre de bits. Il est également possible de considérer globalement l'ensemble de ces deux opérations, mais les techniques de calcul correspondantes sont plus compliquées [8].

5.11 LIMITATION DU NOMBRE DE BITS DES MÉMOIRES INTERNES

La limitation du nombre de bits des mémoires dans un filtre constitue une source de dégradations du signal à la traversée de ce filtre.

Dans un filtre RIF il est possible d'éviter cette dégradation. En effet si b_d désigne le nombre de bits des données d'entrée, b_c étant celui des coefficients, il suffit de pouvoir faire l'accumulation des produits à $(b_d + b_c)$ bits pour réaliser exactement les calculs définis par (5.38). De nombreux circuits multiplieurs et processeurs permettent cette opération. Si les produits sont arrondis à b_m bits pour simplifier les circuits de multiplication et accumulation; il apparaît un bruit appelé bruit de calcul.

Dans les structures considérées au paragraphe précédent ce bruit s'ajoute en sortie de filtre.

La figure 5.22 montre le cadrage des nombres dans le filtre.

Compte tenu des valeurs des coefficients h_i, les produits sont décalés d'un nombre de bits b_0 qui, pour un passe-bas, s'écrit d'après (5.45) :

$$b_0 = \log 2 \left(\frac{f_e}{f_1 + f_2} \right)$$

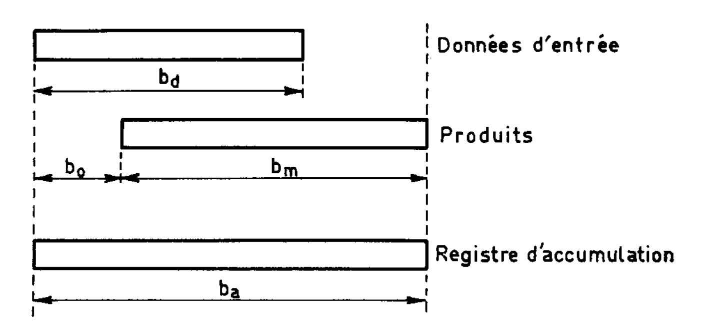


FIG. 5.22. *Cadrage des nombres dans un filtre RIF.*

La sortie du filtre est obtenue dans un accumulateur ayant au moins $b_a = (b_0 + b_m)$ bits.

En fait le nombre de bits b_i à l'intérieur de la machine doit être supérieur à cette valeur b_a pour éviter les débordements, bien qu'avec une représentation en complément à deux des débordements temporaires soient acceptables. Ce nombre de bits b_i va maintenant être relié aux spécifications du filtre.

Au signal qui se présente à l'entrée du filtre est superposé un bruit dont la puissance est désignée par B_1. Cette puissance de bruit est généralement liée au nombre de bits b_d utilisés pour représenter le signal; en fait elle est égale à k_0 fois, avec $k_0 \geqslant 1$, la puissance de bruit engendrée par la quantification à b_d bits. Si le bruit de calcul a pour puissance B_c, si le signal d'entrée S et le bruit superposé B_1 ont une distribution spectrale uniforme, le rapport signal à bruit SB en sortie du filtre s'exprime en décibels (dB) par :

$$SB = 10 \log \left[\frac{S.\dfrac{f_1 + f_2}{f_e}}{B_1 \dfrac{f_1 + f_2}{f_e} + B_c} \right] \tag{5.49}$$

Alors, la réduction ΔSB du rapport signal à bruit à la traversée du filtre s'écrit :

$$\Delta SB = 10 \log \left[1 + \frac{B_c}{B_1} \frac{f_e}{f_1 + f_2} \right] \tag{5.50}$$

Si cette dégradation doit rester faible, il vient :

$$\Delta SB \simeq 4{,}3 . \frac{B_c}{B_1} \frac{f_e}{f_1 + f_2} \tag{5.51}$$

Il faut maintenant déterminer la relation entre la puissance du bruit de calcul B_c et le nombre de bits des mémoires internes b_i.

En prenant comme unité la valeur maximale des nombres d'entrée $x(n)$, c'est-à-dire :

$$|x(n)| \leqslant 1$$

on a en sortie, d'après la relation de définition (5.38) :

$$|y(n)| \leqslant \sum_{i=0}^{N-1} |a_i|$$

Pour un filtre passe-bas de réponse unitaire à la fréquence zéro, on a :

$$\sum_{i=0}^{N-1} a_i = 1 \qquad (5.52)$$

Dans ces conditions la somme des valeurs absolues des coefficients reste généralement inférieure à quelques unités et on peut considérer que les inégalités suivantes sont vérifiées :

$$1 \leqslant \sum_{i=0}^{N-1} |a_i| < 2 \qquad (5.53)$$

L'arrondi à b_i bits dans le processus d'accumulation amène alors un bruit de calcul B_c tel que :

$$B_c = N \, . \, \frac{2^{2(2-b_i)}}{12}$$

alors que le bruit à l'entrée B_1 a pour valeur :

$$B_1 = k_0 \, . \, \frac{2^{2(2-b_d)}}{12} \quad \text{avec} \quad k_0 \geqslant 1$$

Avec la même approximation de N qu'au paragraphe précédent, la dégradation du rapport signal à bruit à la traversée du filtre s'écrit :

$$\Delta SB \simeq 4{,}3 \, . \, \frac{3f_e}{\Delta f} \cdot \frac{4}{k_0 \, 2^{2(b_i - b_d)}} \cdot \frac{f_e}{f_1 + f_2} \qquad (5.54)$$

En général les spécifications du filtre imposent une limite à la valeur ΔSB. En supposant $k_0 = 1$ une estimation du nombre de bits des mémoires dans la machine est donnée par :

$$b_i \simeq b_d + 3 + \frac{1}{2}\left[\log 2\left(\frac{1}{\Delta SB}\right) + \log 2\left(\frac{f_e}{\Delta f}\right) + \log 2\left(\frac{f_e}{f_1 + f_2}\right)\right] \qquad (5.55)$$

Il faut bien noter que la validité de cette estimation est limitée aux faibles valeurs du terme ΔSB, exprimé en décibels. Il apparaît que les filtres à bande passante étroite demandent plus de bits ; en fait il est possible de procéder à un recadrage interne des nombres, en tenant compte de la réduction de puissance du signal après filtrage, qui correspond au nombre de bits b_R avec :

$$b_R = \frac{1}{2} \log 2\left(\frac{f_e}{f_1 + f_2}\right)$$

Avec recadrage, le nombre de bits des mémoires dans la machine est donné par $b_{iR} = b_i - b_R$, c'est-à-dire :

$$b_{iR} \simeq b_d + 3 + \frac{1}{2} \log 2 \left(\frac{1}{\Delta SB} \right) + \frac{1}{2} \log 2 \left(\frac{f_e}{\Delta f} \right) \quad (5.56)$$

Les estimations données dans ce paragraphe et les précédents fournissent une évaluation de la complexité des machines nécessaire pour réaliser les fonctions de filtrage RIF.

5.12 FONCTION DE TRANSFERT EN Z D'UN FILTRE RIF

La fonction de transfert en Z d'un filtre RIF à N coefficients est un polynôme de degré N – 1 qui s'écrit (5.7) :

$$H(Z) = \sum_{i=0}^{N-1} a_i Z^{-i}$$

Ce polynôme possède N – 1 racines Z_i ($1 \leqslant i \leqslant N-1$) dans le plan complexe et s'écrit sous la forme d'un produit de facteurs :

$$H(Z) = a_0 \prod_{i=1}^{N-1} (1 - Z_i Z^{-1}) \quad (5.57)$$

Ces racines possèdent des particularités en raison des propriétés des filtres RIF.

D'abord si les coefficients sont réels, à toute racine complexe Z_i correspond une racine complexe conjuguée $\overline{Z}_i$, de sorte que H(Z) s'écrit sous forme d'un produit de termes du premier degré et de termes du second degré à coefficients réels. Un terme du second degré s'écrit ainsi :

$$H_2(Z) = 1 - 2Re(Z_i) Z^{-1} + |Z_i|^2 Z^{-2} \quad (5.58)$$

D'autre part, la symétrie des coefficients d'un filtre à phase linéaire doit apparaître dans la décomposition en produits de facteurs. Pour un terme du second degré à coefficients réels il faut, si les racines sont complexes, que $|Z_i| = 1$, c'est-à-dire que le zéro soit sur le cercle unité. Pour un terme du 4e degré à coefficients réels, il faut que les 4 racines complexes soient les suivantes : Z_i, $\overline{Z}_i$, $\frac{1}{Z_i}$, $\frac{1}{\overline{Z}_i}$; c'est-à-dire :

$$H_4(Z) = 1 - 2Re\left(Z_i + \frac{1}{Z_i}\right)Z^{-1} + \left[|Z_i|^2 + \frac{1}{|Z_i|^2} + 4Re\,(Z_i)\,Re\left(\frac{1}{Z_i}\right)\right]Z^{-2}$$
$$- 2Re\left(Z_i + \frac{1}{Z_i}\right)Z^{-3} + Z^{-4} \qquad (5.59)$$

Dans ces conditions un filtre RIF à phase linéaire peut se décomposer en un ensemble de filtres élémentaires du deuxième ou quatrième degré ayant la propriété de symétrie des coefficients.

Les racines du filtre passe-bas à 15 coefficients donné comme exemple dans le paragraphe 5.7 ont été calculées. Les affixes des 14 zéros sont les suivants :

$Z_1 = -0{,}976 \pm j\,0{,}217 \quad Z_5 = 0{,}492 \pm j\,0{,}266\,;$
$Z_2 = -0{,}797 \pm j\,0{,}603 \quad Z_6 = 1{,}573 \pm j\,0{,}851\,;$
$Z_3 = -0{,}512 \pm j\,0{,}859 \quad Z_7 = 0{,}165$
$Z_4 = -0{,}271 \pm j\,0{,}962 \quad Z_8 = 6{,}052$

Leur position dans le plan complexe est donné sur la figure 5.23.

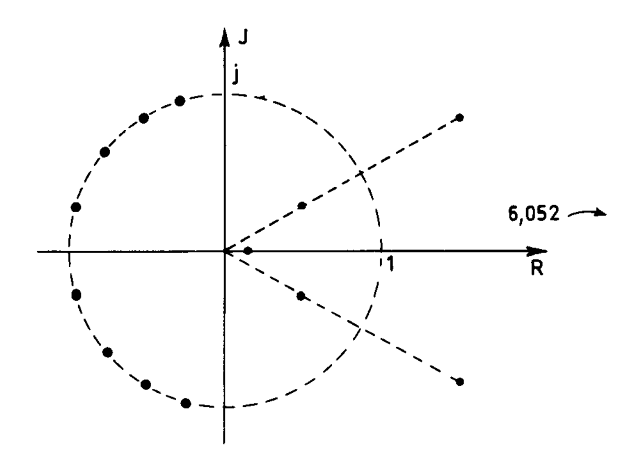


FIG. 5.23. *Configuration des zéros d'un filtre RIF*

Cette figure illustre les caractéristiques de la réponse en fréquence du filtre et est à rapprocher de la figure 5.16. Les couples de racines caractéristiques de la linéarité en phase apparaissent également. Si cette contrainte n'est plus imposée la configuration des racines est modifiée.

Il est intéressant d'observer que, comme le montre l'expression (5.14), la réponse en fréquence d'un filtre à phase linéaire à nombre de coefficients pair s'annule à la demi-fréquence d'échantillonnage $f_e/2$. Un tel filtre possède donc un zéro à $f_e/2$. De même, si un filtre à nombre de coefficients impair possède un zéro à $f_e/2$, ce zéro est double, ce qui assure la symétrie de la réponse en fréquence au voisinage de $f_e/2$.

5.13 FILTRES À DÉPHASAGE MINIMAL

Le temps de propagation à travers un filtre à phase linéaire peut être trop important pour certaines applications. D'autre part, il n'est pas toujours possible ou intéressant d'utiliser la symétrie des coefficients d'un filtre à phase linéaire pour simplifier les calculs [9]. Alors si la linéarité en phase n'est pas une caractéristique imposée, on peut espérer réduire la complexité du filtre en abandonnant cette contrainte. En effet une fonction de transfert à phase linéaire peut être considérée comme le produit d'une fonction à déphasage minimal par une fonction de déphaseur pur. La condition pour qu'une fonction de transfert en Z soit à déphasage minimal est que ses racines soient à l'intérieur ou sur le cercle unité. Ce point est développé au chapitre 9.

On peut obtenir les coefficients d'un filtre à déphasage minimal à partir des coefficients d'un filtre à phase linéaire optimal d'une manière simple.

En effet soit un filtre à phase linéaire à N = 2P + 1 coefficients dont la réponse en fréquence s'écrit :

$$H(f) = e^{-j2\pi fPT}\left[h_0 + 2\sum_{i=1}^{P} h_i \cos(2\pi f i T)\right]$$

Les ondulations en bandes passante et affaiblie sont δ_1 et δ_2 respectivement. Examinons le filtre obtenu en ajoutant δ_2 à la réponse précédente et en recadrant pour approcher l'unité en bande passante. Sa réponse $H_2(f)$ est telle que :

$$H_2(f) = e^{-j2\pi fPT}\cdot\frac{1}{1+\delta_2}\left[h_0 + \delta_2 + 2\sum_{i=1}^{P} h_i \cos(2\pi f i T)\right] \qquad (5.60)$$

En bande passante il présente des ondulations d'amplitude δ_1' telle que :

$$\delta_1' = \frac{\delta_1}{(1+\delta_2)}$$

Sa réponse en bande affaiblie est représentée sur la figure 5.24 ; les ondulations sont limitées à $\delta_2' = \dfrac{2\delta_2}{(1+\delta_2)}$.

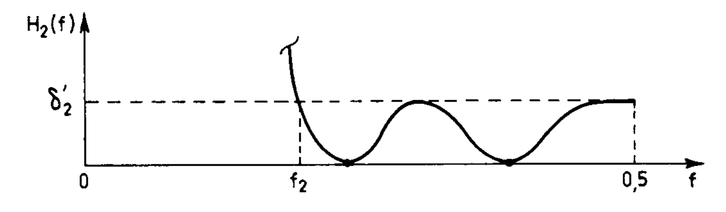


FIG. 5.24. *Ondulations en bande affaiblie du filtre surélevé*

Ce filtre est à phase linéaire car la symétrie des coefficients est conservée. Par contre on peut observer que les zéros de la fonction de transfert en Z qui sont sur le cercle unité sont doubles car $H_2(f)$ ne devient pas négatif. Dans ces conditions la configuration des zéros est celle de la figure 5.25.

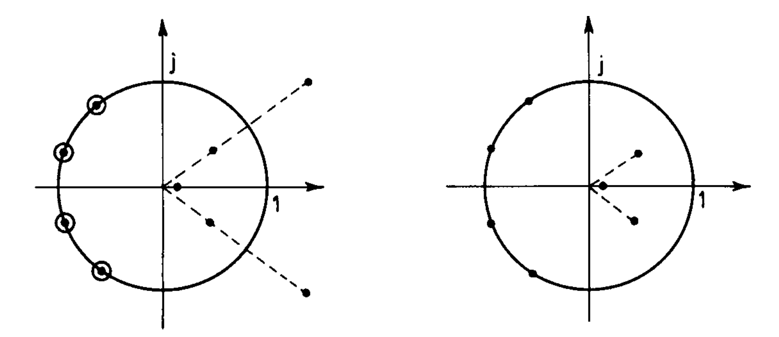


FIG. 5.25. *Configuration des zéros de* $H_2(f)$ *et du filtre à déphasage minimal*

Les zéros qui ne sont pas sur le cercle unité ne sont pas doubles. Cependant le module de la fonction $H_2(f)$ n'est pas modifié, à une constante près, si l'on remplace les zéros Z_i extérieurs au cercle unité par les zéros $\frac{1}{Z_i}$, qui sont intérieurs au cercle unité et deviennent alors doubles également. En effet cette opération revient simplement à une multiplication par G(Z) telle que :

$$G(Z) = \frac{\left(1 - \frac{Z^{-1}}{Z_i}\right)\left(1 - \frac{Z^{-1}}{\overline{Z}_i}\right)}{(1 - Z_i Z^{-1})(1 - \overline{Z}_i Z^{-1})}$$

Or, sur le cercle unité $Z^{-1} = \overline{Z}$ et la symétrie par rapport à l'axe réel conduit à l'égalité :

$$\left| \frac{(Z^{-1} - Z_i)(Z^{-1} - \overline{Z}_i)}{(Z - Z_i)(Z - \overline{Z}_i)} \right|_{Z = e^{j\omega}} = 1 \qquad (5.61)$$

Par suite :

$$|G(e^{j2\pi f})| = \frac{1}{|Z_i|^2}$$

Dans ces conditions on peut écrire :

$$H_2(f) = H_m^2(f) . K \quad (K : \text{constante})$$

où $H_m(f)$ est la réponse d'un filtre qui a une fonction de transfert en Z dont les P zéros sont simples et à l'intérieur ou sur le cercle unité. Ce filtre satisfait à la condi-

tion de phase minimale, il possède P + 1 coefficients et les amplitudes des ondulations en bandes passante et affaiblie sont δ_{m_1} et δ_{m_2} telles que :

$$\delta_{m_1} = \sqrt{1 + \frac{\delta_1}{1+\delta_2}} - 1 \simeq \frac{1}{2}\,\frac{\delta_1}{1+\delta_2} \simeq \frac{\delta_1}{2} \tag{5.62}$$

$$\delta_{m_2} = \sqrt{\frac{2\delta_2}{1+\delta_2}} \simeq \sqrt{2\delta_2} \tag{5.63}$$

Pour calculer ce filtre il suffit de partir du filtre à phase linéaire dont les paramètres δ_1 et δ_2 sont déterminés à partir de δ_{m_1} et δ_{m_2} et de suivre la procédure qui a été décrite. Un inconvénient de cette procédure est qu'elle exige l'extraction des racines d'un polynôme de degré N – 1, ce qui limite les valeurs de N envisageables. D'autres procédures peuvent être utilisées [10], [11]. En particulier une méthode simple peut être déduite de la prédiction linéaire, comme indiqué au chapitre 13.

Une estimation N'_e de l'ordre du filtre RIF à déphasage minimal peut être déduite de la relation (5.32). Selon la procédure décrite ci-dessus, pour des spécifications δ_1 et δ_2, il vient :

$$N'_e \simeq \frac{1}{2}\cdot\frac{2}{3}\,.\log\left[\frac{1}{10\,.\,\delta_1\,.\,\delta_2^2}\right].\,\frac{f_e}{\Delta f}$$

ou encore :

$$N'_e \simeq N_e - \frac{1}{3}\,.\log\left[\frac{1}{10\,.\,\delta_1}\right].\,\frac{f_e}{\Delta f} \tag{5.64}$$

La validité de cette formule est naturellement limitée au cas où $\delta_1 \ll 0{,}1$. Le gain obtenu sur l'ordre du filtre avec le déphasage minimal est fonction de l'ondulation en bande passante ; il reste relativement modeste en général.

Exemple

Soit le gabarit de filtre passe-bas suivant (exemple 3, paragraphe 5.6) :

$$\delta_{m_1} = 0{,}0411\,;\quad \delta_{m_2} = 0{,}0137\,;\quad \Delta f = 0{,}115$$

Alors il vient :

$$\delta_1 = 0{,}0822.\quad \delta_2 = 0{,}0000938$$

et le nombre de coefficients nécessaire pour le filtre à phase linéaire correspondant est estimé à : $N_e = 24$, ce qui conduit à $N'_e = 12$. En réalité, on vérifie que le filtre à déphasage minimal satisfaisant au gabarit nécessite 11 coefficients au lieu de 15 pour le filtre à phase linéaire.

En conclusion, lorsque les symétries apportées par la linéarité en phase ne peuvent pas être exploitées, il peut être avantageux de recourir aux filtres à déphasage minimal.

5.14 CALCUL DES FILTRES À TRÈS GRAND NOMBRE DE COEFFICIENTS

Quand le nombre de coefficients du filtre est très élevé, par exemple un millier ou plus, ce qui correspond à des bandes de transition extrêmement faibles, de l'ordre de quelques millièmes, les techniques d'optimisation deviennent difficiles à utiliser ou ne convergent plus. On peut alors utiliser des techniques sous-optimales mais qui ne nécessitent que le calcul de filtres à nombre de coefficients réduit. C'est le cas de la méthode dite du masquage en fréquence [12].

Soit à réaliser un filtre H(Z) dont la bande de transition Δf est centrée sur la fréquence de coupure f_c. On commence par calculer un filtre passe-bas $H_0(Z^M)$, avec une fréquence d'échantillonnage réduite à f_e/M, avec $M < \dfrac{f_e}{4\Delta f}$, et tel que la bande de transition d'une des répliques de ce filtre sur l'axe des fréquences coïncide avec la bande de transition du filtre désiré, comme le montre la figure 5.26*b*.

Ensuite, on construit à partir de $H_0(Z^M)$ deux filtres complémentaires comme indiqué sur la figure 5.26*b*, ce qui nécessite pour $H_0(Z^M)$ un nombre impair de coefficients, 2P + 1.

On obtient un diagramme à 2 branches, auxquelles on applique les filtres $G_1(Z)$ et $G_2(Z)$, dits interpolateurs et ayant les réponses données à la figure 5.26*c*. Il suffit alors de sommer les sorties pour obtenir le filtre désiré de la figure 5.26*a*. Le schéma global est celui de la figure 5.27.

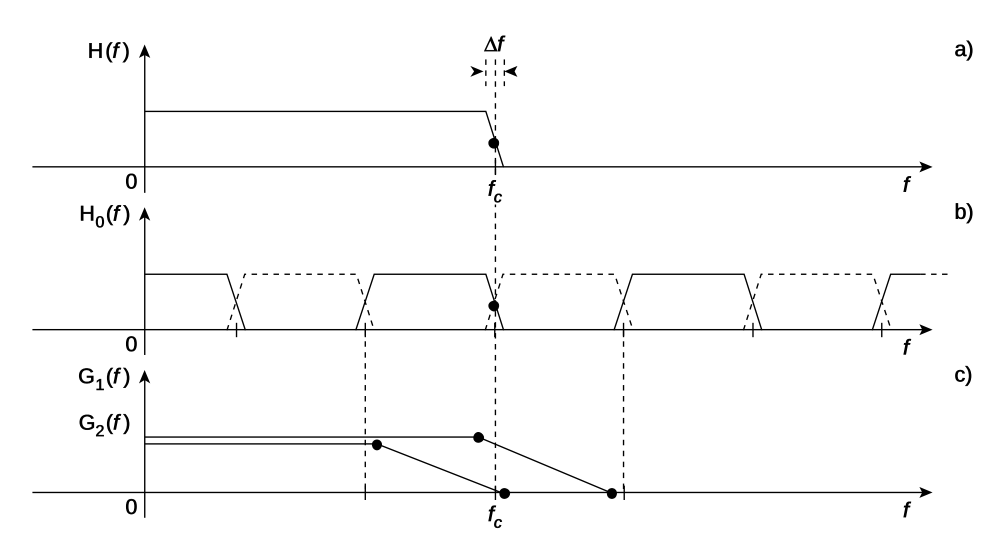


FIG. 5.26. *Principe du masquage en fréquence*
a) Filtre désiré
b) Filtre sous-échantillonné
c) Filtres interpolateurs

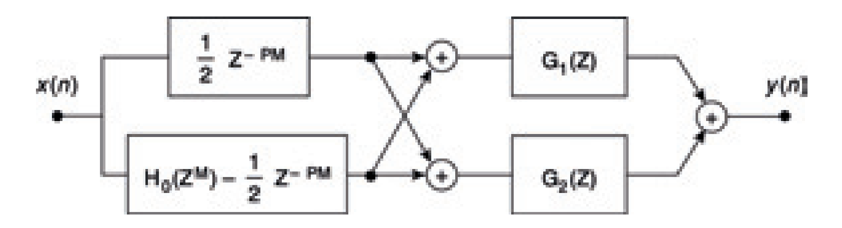


FIG. 5.27. *Schéma du filtre selon la technique de masquage en fréquence*

La procédure nécessite ainsi le calcul de 3 filtres ayant comme bandes de transition $M\Delta f$, $f_c - k\,\frac{f_e}{M}$, $(k+1)\,\frac{f_e}{M} - f_c$, où k est l'entier qui permet d'encadrer la fréquence de coupure f_c.

La fonction de transfert H(Z) du filtre désiré prend la forme :

$$H(Z) = H_0(Z^M)\,G_1(Z) + [Z^{-PM} - H_0(Z^M)]\,G_2(Z) \qquad (5.65)$$

ce qui fournit les valeurs des coefficients.

A noter que le schéma de la figure 5.27 fournit une réalisation efficace du filtre global puisque le filtre $H_0(Z^M)$ a M – 1 coefficients nuls entre deux coefficients non nuls. Ce schéma peut se simplifier comme indiqué sur la figure 5.28. On peut prendre comme filtres interpolateurs $F_1(Z) = G_1(Z) + G_2(Z)$ et $F_2(Z) = G_1(Z) - G_2(Z)$, mais ces filtres peuvent aussi se calculer directement à partir de leurs spécifications déduites de la figure 5.26.

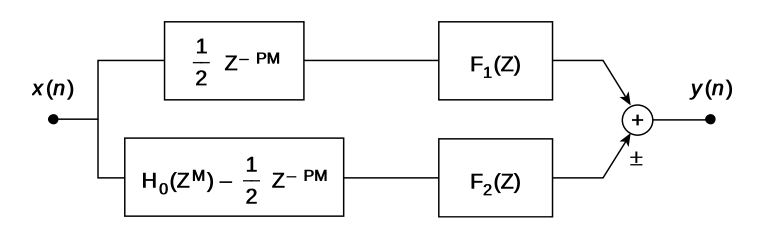


FIG. 5.28. *Schéma simplifié du filtre par masquage en fréquence*

5.15 FILTRES RIF À DEUX DIMENSIONS

Un filtre RIF à deux dimensions est défini par une relation entre la sortie $y(n, m)$ et l'entrée $x(n, m)$ qui s'écrit :

$$y(n, m) = \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} a_{ij}\,x(n-i, m-j) \qquad (5.66)$$

L'ensemble des coefficients a_{ij} constitue une matrice $A_{N_1N_2}$ de dimension $N_1 \times N_2$. La fonction de transfert à 2 variables correspondante, $H(Z_1, Z_2)$ s'exprime en fonction de cette matrice par :

$$H(Z_1, Z_2) = \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} a_{ij}\, Z_1^{-i}\, Z_2^{-j} \tag{5.67}$$

ou encore, sous forme vectorielle :

$$H(Z_1, Z_2) = [1, Z_1^{-1}, \ldots, Z_1^{-(N_1-1)}]\, A_{N_1N_2} \begin{bmatrix} 1 \\ Z_2^{-1} \\ \vdots \\ Z_2^{-(N_2-1)} \end{bmatrix} \tag{5.68}$$

La matrice des coefficients $A_{N_1N_2}$ est aussi appelée le masque. A titre d'exemple, les filtres passe-haut suivants sont d'utilisation courante en traitement d'image :

$$A' = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} ; \quad A'' = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ -1 & -1 & -1 \end{bmatrix}$$

Le filtre A′ est dit de Sobel et A″ de Prewitt. Dans les procédures d'extraction des contours dans une image, ils sont utilisés deux fois, comme ci-dessus et après rotation de 90°.

Les coefficients des filtres à deux dimensions peuvent être calculés directement à partir des spécifications dans le domaine des fréquences à deux dimensions.

Quand la réponse impulsionnelle est une fonction paire par rapport aux deux variables, la réponse en fréquence et les coefficients peuvent être obtenus à partir d'un filtre à une dimension et à phase linéaire. En effet soit $H(\omega)$ la réponse en fréquence d'un tel filtre, qui, d'après (5.13) en négligeant le terme de phase, s'exprime par :

$$H(\omega) = h_0 + 2 \sum_{i=1}^{P} h_i \cos i\omega$$

Or, il existe entre $\cos i\omega$ et $\cos \omega$ une relation polynomiale :

$$\cos i\omega = T_i (\cos \omega) \tag{5.69}$$

où $T_i(x)$ est le polynôme de Tchebycheff de degré i. Dans ces conditions $H(\omega)$ s'écrit aussi :

$$H(\omega) = \sum_{i=0}^{P} g_i (\cos \omega)^i \tag{5.70}$$

Ensuite, le changement de variables :

$$\cos \omega = H_1(\omega_1, \omega_2) = \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} t(k, l) \cos k\omega_1 \cos l\omega_2 \tag{5.71}$$

conduit à la fonction à deux variables suivante :

$$H(e^{\omega_1}, e^{\omega_2}) = \sum_{i=0}^{P} g_i \left(\sum_{k=0}^{K-1} \sum_{l=0}^{L-1} t(k, l) \cos k\omega_1 \cos l\omega_2 \right) \quad (5.72)$$

qui peut être réécrite sous la forme :

$$H(\omega_1, \omega_2) = \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} h_{ij} \cos i\omega_1 \cos j\omega_2 \quad (5.73)$$

avec :

$$N_1 = 2KP + 1 ; \quad N_2 = 2LP + 1$$

La fonction $t(k, l)$ peut être choisie pour qu'à chaque valeur de ω corresponde un contour dans le plan (ω_1, ω_2). Par exemple pour :

$$\cos \omega = \frac{1}{2} [\cos \omega_1 + \cos \omega_2 + \cos \omega_1 \cos \omega_2 - 1] \quad (5.74)$$

on obtient approximativement une symétrie circulaire, comme le montre le développement limité de $\cos \omega_1$ pour ω_1 petit.

La figure 5.29 montre un exemple de réponse de filtre calculé par cette méthode.

La réalisation d'un filtre à deux dimensions peut se faire par application directe de la définition (5.66). Dans le cas des filtres déduits d'une fonction monodimensionnelle, la réalisation peut être simplifiée en utilisant la relation (5.72) et en procédant comme pour un filtre à une dimension et P + 1 coefficients $g_i (0 \leqslant i \leqslant P)$, mais dans lequel le retard est remplacé par la cellule à deux dimensions correspondant à la fonction $H_1(\omega_1, \omega_2)$ [13].

FIG. 5.29. *Filtre RIF à deux dimensions calculé à partir d'un filtre 1-D à phase linéaire*

Un cas de réalisation particulièrement simple est celui des filtres dits séparables, pour lesquels la matrice des coefficients est dyadique, c'est-à-dire :

$$A_{N_1 N_2} = V_1 V_2^t$$

où V_1 et V_2 sont des vecteurs. Alors, conformément à la relation (5.68), la fonction de transfert se factorise :

$$H(z_1, z_2) = H_1(z_1)\, H_2(z_2). \tag{5.75}$$

Les spécifications de tels filtres sont soumises à des limitations. D'abord, elles doivent correspondre à la symétrie quadrantale suivant les axes de coordonnées. Comme indiqué sur la figure 5.30, le domaine des fréquences utiles se divise en quatre parties : passe-bas/passe-bas (BB), passe-bas/passe-haut (BH), passe-haut/passe-bas (HB) et passe-haut/passe-haut (HH).

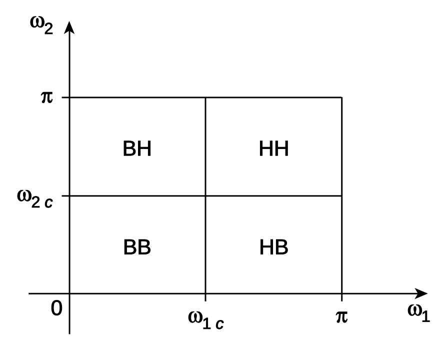


Fig. 5.30. *Domaines de fréquence pour un filtre 2D séparable*

Ensuite, les spécifications d'ondulation doivent être définies en conséquence. Par exemple, pour un filtre 2D de type passe-bas, le domaine HH subit l'affaiblissement des 2 filtres, horizontal et vertical. Une illustration est donnée par la figure 5.31 qui montre la réponse en fréquence d'un filtre 2D séparable, basé sur le filtre demi-bande de la figure 5.13.

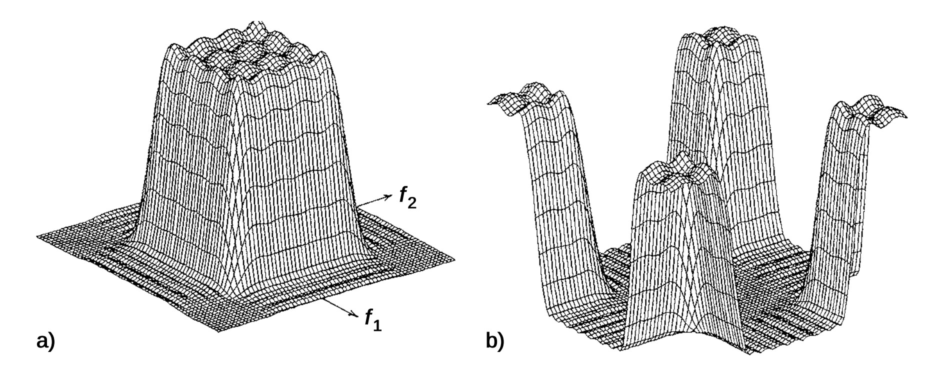


Fig. 5.31. *Filtre 2D demi-bande séparable*

a) Passe-bas/Passe-bas
b) Passe-haut/Passe-haut

La réalisation peut se faire suivant la définition, c'est-à-dire qu'un tableau de données représentant une image peut être traité ligne par ligne avec le filtre horizontal et colonne par colonne avec le filtre vertical.

Quand l'image est soumise à un balayage horizontal comme en télévision, le signal apparaît en fait comme mono-dimensionnel et peut être traité comme tel. Si chaque ligne comporte N points, la fonction de transfert s'écrit :

$$H(z_1, z_2) = H_1(z)\, H_2(z^N) \tag{5.76}$$

Par exemple, pour le filtre de Sobel A′, on a :

$$A' = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} . [-1 \quad 0 \quad 1]$$

et le circuit correspondant est donné à la figure 5.32. La réalisation est particulièrement simple, les circuits ne comportant pas de multiplieurs.

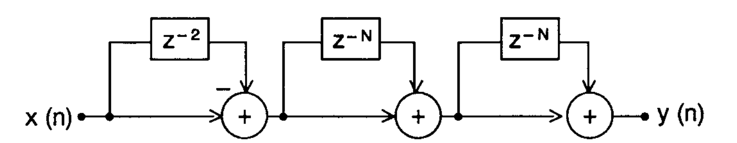


Fig. 5.32. *Réalisation d'un filtre d'extraction des contours*

5.16 CALCUL DES COEFFICIENTS DE FILTRES RIF-2D PAR LA MÉTHODE DES MOINDRES CARRÉS

La méthode va être développée pour un cas particulier important, celui des filtres à symétrie quadrantale. Deux types de filtres correspondent à cette catégorie, les filtres en rectangle et les filtres en losange, avec les domaines de fréquence de la figure 5.33.

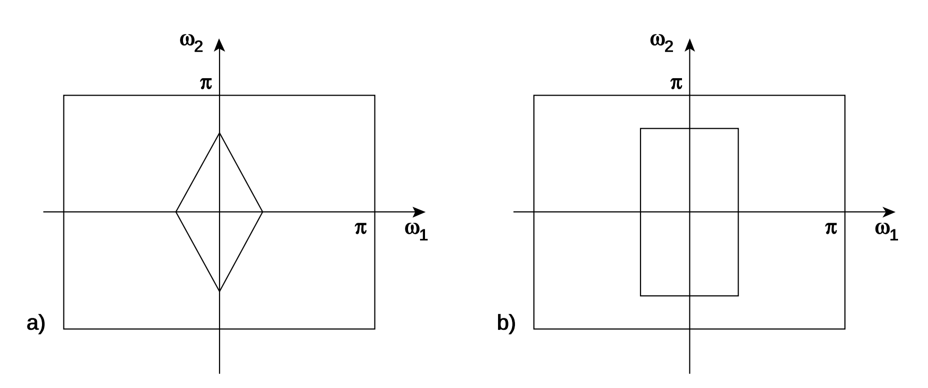


Fig. 5.33. *Filtre 2D en losange (a) et en rectangle (b)*

La réponse en fréquence d'un filtre à phase nulle ayant (2M + 1) × (2N + 1) coefficients avec symétrie quadrantale s'exprime par :

$$H(\omega_1, \omega_2) = h_{00} + 2 \sum_{i=1}^{M} h_{i0} \cos i\omega_1$$

$$+ 2 \sum_{i=1}^{N} h_{0j} \cos j\omega_2 + 4 \sum_{i=1}^{M} \sum_{j=1}^{M} h_{ij} \cos i\omega_1 \cos j\omega_2 \tag{5.77}$$

Au total le filtre possède (1 + M + N + MN) coefficients h_{ij} de valeurs différentes.

La méthode des moindres carrés avec pondération va être appliquée directement, pour approcher la réponse désirée, $D(\omega_1, \omega_2)$. Avec un facteur de suréchantillonnage égal à k, la fonction quadratique d'écart, ou fonction coût, à minimiser s'écrit :

$$J = \sum_{m=0}^{K_M} \sum_{n=0}^{K_N} \left| H\left(\frac{m\pi}{K_M}, \frac{n\pi}{K_N}\right) - D\left(\frac{m\pi}{K_M}, \frac{n\pi}{K_N}\right) \right|^2 W\left(\frac{m\pi}{K_M}, \frac{n\pi}{K_N}\right) \tag{5.78}$$

avec $K_M = k(M + 0{,}5)$ et $K_N = k(N + 0{,}5)$ afin de couvrir la totalité du domaine des fréquences utiles.

La fonction de pondération $W(\omega_1, \omega_2)$ permet d'ajuster l'approximation en fonction des spécifications d'ondulation par exemple.

Avec des notations simplifiées, il vient :

$$J = \sum_{m=0}^{K_M} \sum_{n=0}^{K_N} e^2(m, n)\, W(m, n) \tag{5.79}$$

Le minimum de la fonction coût est obtenu pour :

$$\sum_{m=0}^{K_M} \sum_{n=0}^{K_N} e(m, n)\, W(m, n) \frac{\partial e(m, n)}{\partial h_{ij}} = 0 \tag{5.80}$$

Ce qui donne un système de (1 + M + N + MN) équations.

En désignant par $[h_{ij}]$ le vecteur des coefficients et par $V(m, n)$ le vecteur fréquentiel :

$$V^t(m, n) = \left[1, \ldots, 2\cos\left(i\frac{m\pi}{K_M}\right), \ldots, 2\cos\left(j\frac{n\pi}{K_N}\right), \ldots,\right.$$

$$\left. 4\cos\left(i\frac{m\pi}{K_M}\right)\cos\left(j\frac{n\pi}{K_N}\right)\ldots\right]$$

la solution s'écrit :

$$[h_{ij}] = \left[\sum_{m=0}^{K_M} \sum_{n=0}^{K_N} W(m, n)\, V(m, n)\, V^t(m, n)\right]^{-1}$$

$$\left[\sum_{m=0}^{K_M} \sum_{n=0}^{K_N} W(m, n)\, V(m, n)\, D(m, n)\right] \tag{5.81}$$

Si le nombre de coefficients est pair, il faut modifier les paramètres. Par exemple, pour un filtre à (2M) × (2N + 1) coefficients, il faut prendre :

$$V^t(m, n) = \left[\ldots, 2\cos\left[(i-0{,}5)\,\frac{m\pi}{K_M}\right], \ldots\right.$$
$$\left. 4\cos\left((i-0{,}5)\,\frac{m\pi}{K_M}\right)\cos\left(j\,\frac{n\pi}{K_N}\right)\ldots\right] \qquad (5.82)$$

avec $K_M = kM$ et $K_N = k(N + 0{,}5)$.

Le vecteur des coefficients obtenu dans ce cas possède (M + MN) éléments.

Une caractéristique importante des filtres utilisés en traitement d'image est la réponse à l'échelon unité. En effet, les suroscillations à la transition peuvent produire des répétitions de contours et ainsi dégrader l'image. En modifiant la réponse désirée $D(\omega_1, \omega_2)$ par une inclinaison à la fin de la bande passante et au début de la bande affaiblie, il est possible de réduire ces suroscillations.

La méthode est illustrée par le calcul d'un filtre rectangulaire avec (2M + 1) × (2N + 1) = 9 × 9 coefficients, avec 0,125 et 0,25 comme fin de bande passante et début de bande affaiblie sur l'axe des fréquences horizontal et 0,0625 et 0,125 sur l'axe vertical. Les 25 coefficients différents obtenus sont donnés par le tableau :

$$h_{ij} = \begin{bmatrix} 0{,}052427 & 0{,}0419028 & 0{,}0184534 & -0{,}0002861 & -0{,}006258 \\ 0{,}0491981 & 0{,}0393451 & 0{,}0173566 & -0{,}0002629 & -0{,}0059292 \\ 0{,}041534 & 0{,}0332908 & 0{,}0147612 & -0{,}000261 & -0{,}005282 \\ 0{,}0299102 & 0{,}0240605 & 0{,}0107414 & -0{,}0002704 & -0{,}0041828 \\ 0{,}0180912 & 0{,}0146366 & 0{,}0065523 & -0{,}0003836 & -0{,}0031209 \end{bmatrix}$$

et la réponse en fréquence correspondante est donnée à la figure 5.34. Visiblement cette réponse est très proche de celle d'un filtre séparable.
Considérant maintenant un filtre en losange avec (2M + 1) × (2N) = 9 × 8 coefficients, une fin de bande passante à 0,125 et un début de bande affaiblie à 0,25 sur les axes horizontal et vertical, le tableau des coefficients suivant est obtenu pour un quadran :

$$h_{ij} = \begin{bmatrix} 0{,}0763835 & 0{,}0680674 & 0{,}0403862 & 0{,}0130039 & 0{,}000071 \\ 0{,}0642979 & 0{,}03951 & 0{,}0217936 & 0{,}0008111 & -0{,}002745 \\ 0{,}0276109 & 0{,}0195655 & 0{,}0068997 & -0{,}0050102 & -0{,}0110481 \\ 0{,}0065124 & 0{,}0011002 & 0{,}0085984 & -0{,}0099831 & -0{,}0073724 \end{bmatrix}$$

La réponse en fréquence est donnée à la figure 5.35. Le calcul a été mené en cherchant à réduire la réponse à l'échelon unité $g(i, j)$ définie par :

$$g(i, j) = \sum_{i_1=-M}^{i} \sum_{j_1=-N}^{j} h(i_1, j_1) \qquad (5.83)$$

Cette réponse est donnée également sur la figure, où elle a été répétée sur les 4 quadrans, pour fournir une vue complète.

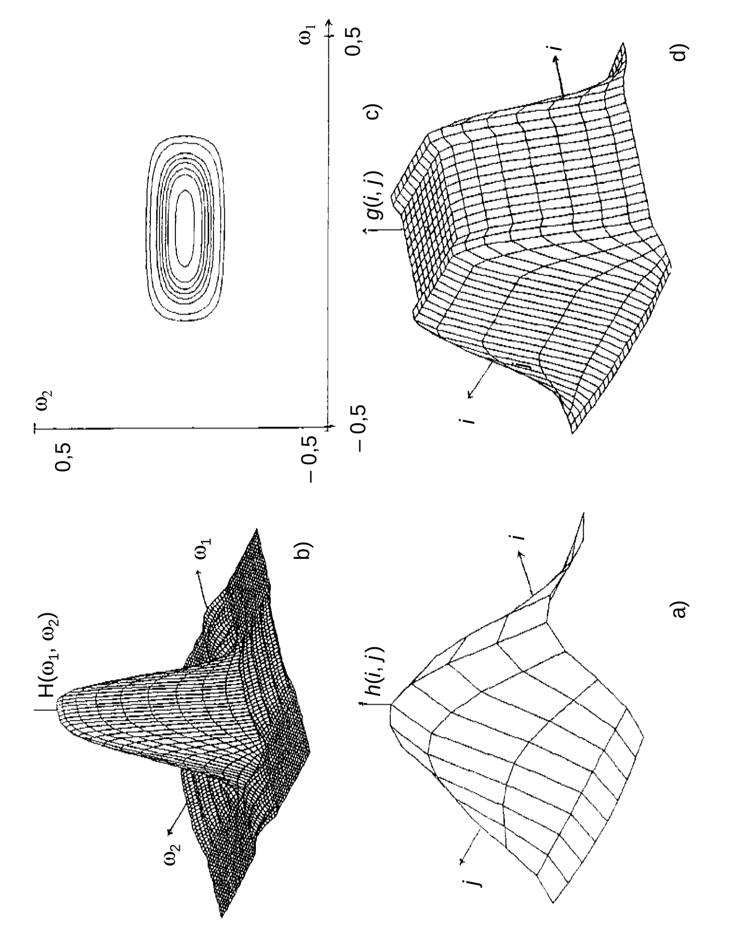


Fig. 5.34. *Filtre rectangulaire à 9 × 9 coefficients*

a) Réponse impulsionnelle
b) Réponse en fréquence
c) Coupe horizontale de la réponse en fréquence
d) Réponse à l'échelon unité

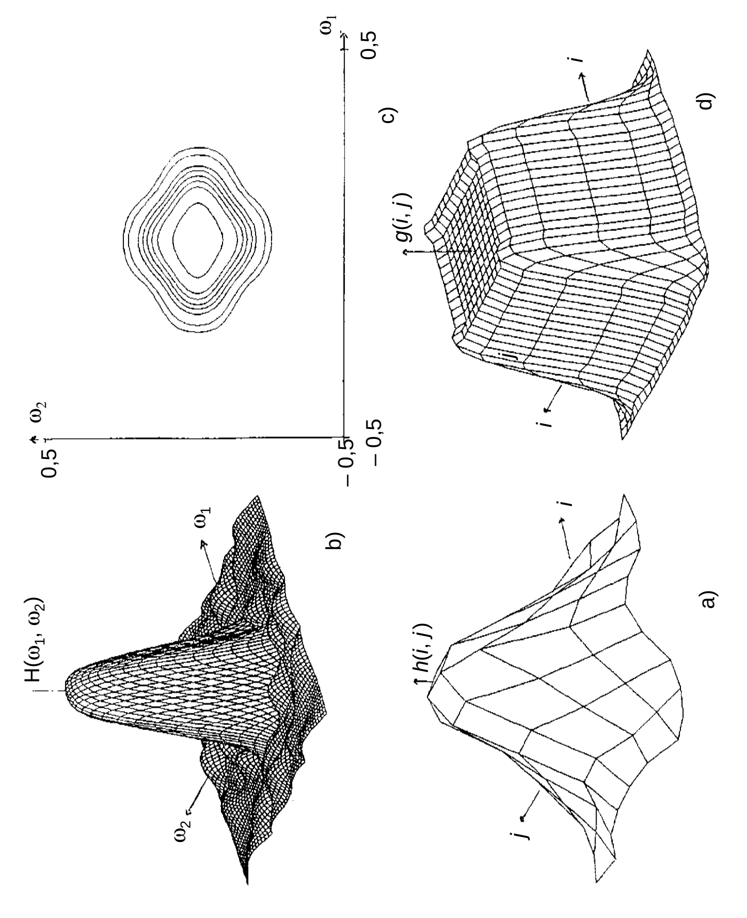


Fig. 5.35. *Filtre en losange à 9 × 8 coefficients*

a) Réponse impulsionnelle
b) Réponse en fréquence
c) Coupe horizontale de la réponse en fréquence
d) Réponse à l'échelon unité

Les deux filtres calculés ont été appliqués à une mire d'évaluation. La figure 5.36 montre l'élimination des répliques obtenue par le filtre rectangulaire et le filtre en losange.

Pour des développements complémentaires sur les techniques de calcul des filtres RIF-2D, y compris avec coefficients en précision limitée et contraintes sur la réponse à l'échelon unité, on peut se reporter aux références [14, 15].

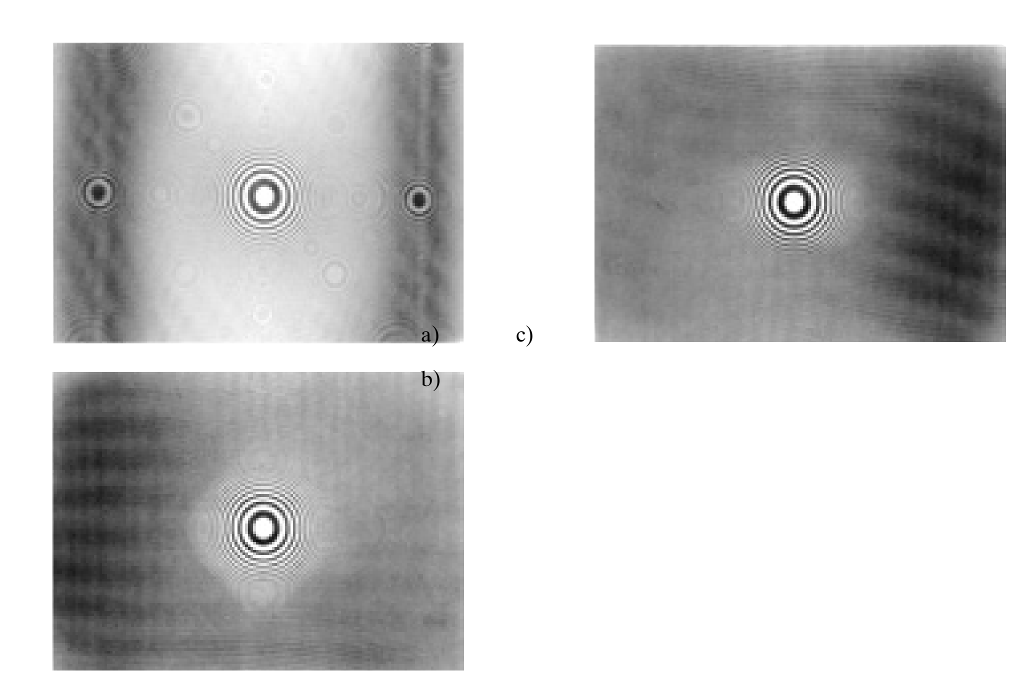


FIG. 5.36. *Filtrage d'une mire d'évaluation*

a) Image d'origine
b) Filtrage en losange de la mire
c) Filtrage rectangulaire de la mire

ANNEXE

Exemple de calcul d'un filtre RIF

```
FILTRE A 2 BANDES DE FREQUENCE


NOMBRE DE COEFF. =  32

  ***** COEFFICIENTS *****

 A  1 =   0.01064416 = A 32
 A  2 =  -0.00703074 = A 31
 A  3 =  -0.02489676 = A 30
 A  4 =  -0.04257757 = A 29
 A  5 =  -0.04550193 = A 28
 A  6 =  -0.02596906 = A 27
 A  7 =   0.00801637 = A 26
 A  8 =   0.03449482 = A 25
 A  9 =   0.03164973 = A 24
 A 10 =  -0.00394829 = A 23
 A 11 =  -0.04885845 = A 22
 A 12 =  -0.06340511 = A 21
 A 13 =  -0.01710309 = A 20
 A 14 =   0.08742009 = A 19
 A 15 =   0.20962396 = A 18
 A 16 =   0.29156661 = A 17


BANDE NUMERO            1         2

LIMITE INF.           0.0000    0.1700
LIMITE SUP.           0.1400    0.5000
VAL. RECH.            1.0000    0.0000
PONDERATION           1.00     10.00


ECART                 0.21175   0.02118
ECART EN DB           1.668   -33.484



FREQUENCES EXTREMALES
0.0000  0.0410  0.0801  0.1172  0.1400  0.1700  0.1798  0.2032
0.2305  0.2618  0.2930  0.3243  0.3555  0.3888  0.4200  0.4513
0.4845
```

BIBLIOGRAPHIE

[1] T. W. Parks and J. H. Mac Clellan – Chebyshev Approximation for Non Recursive Digital Filters with Linear Phase. *IEEE Trans. Circuit Theory*, Vol. CT 19, March 1972.

[2] J. H. Mac Clellan, T. W. Parks and L. Rabiner – A Computer Program for Designing Optimum FIR Filters. *IEEE Trans. On Audio and Electroacoustics*, Dec. 1973.

[3] J. Shen and G. Strang – « The Asymptotics of optimal (Equiripple) Filters », *IEEE Trans*, Vol. SP-47, N° 4, April 1999, pp. 1087-1098.

[4] R. Crochière and A. Oppenheim – Analysis of Linear Digital Networks. *Proceedings IEEE,* April 1975.

[5] W. Schussler – On Structures for Non Recursive Digital Filters. Arch. Elek. Ubertragung, June 1972.

[6] M. Bellanger and G. Bonnerot – Premultiplication Scheme for Digital FIR Filters. *IEEE Trans.* Vol. ASSP 26, Feb. 1978.

[7] D. Chan and L. Rabiner – Analysis of Quantization Errors in the Direct Form of FIR Filters. *IEEE Trans on Audio and Electroacoustics*, August 1973.

[8] F. Grenez – Synthèse des filtres numériques non récursifs à coefficients quantifiés. *Annales des Télécom.*, 34, N°s 1-2, 1979.

[9] M. Feldmann, J. Henaff, B. Lacroix and J. C. Rebourg – Design of Minimum Phase Charge-Transfer Transversal Filters. *Electronic Letters*, Vol. 15, N° 8, April 1979.

[10] R. Boite and H. Leich – A new Procedure for the design of High Order minimum Phase FIR Filters. Signal Processing. Vol. 3, N° 2, April 1981, pp. 101-108.

[11] Y. Kamp and C. J. Wellekens – Optimal Design of Minimum-Phase FIR Filters. *IEEE Trans.* Vol. ASSP-31, N° 4, August 1983.

[12] Y. C. Lim and Y. Lian – « The optimum design of one and two-dimensional FIR filters using the frequency response masking technique », *IEEE Trans. On Circuits and Systems-II*, Vol. 40, N° 2, pp. 88-95, Feb. 1993.

[13] D. Dudgeon, R. Mersereau – « Multidimensional Digital Signal Processing », Prentice-Hall Englewood Cliffs, N. J., 1984.

[14] P. Siohan – « Contribution à l'étude des méthodes de conception des filtres numériques RIF : application au traitement d'images ». Thèse de doctorat de l'ENST, mars 1989.

[15] V. Ouvrard et P. Siohan – « Design of 2D video filters with spatial Constraints », Proceedings of EUSIPCO-92, North Holland, Bruxelles, Aug. 1992, pp. 1001-1004.

EXERCICES

1 On considère les 17 premiers coefficients d'un filtre passe-bas de fréquence de coupure égale à 0,25 f_e donnés à la figure V.12. Combien prennent des valeurs différentes ? Donner l'expression de la réponse en fréquence H(f). Rechercher les points de l'axe des

fréquences où elle s'annule et donner l'ondulation maximale. Calculer les zéros de la fonction de transfert en Z du filtre.

2 Soit un filtre dont la fréquence d'échantillonnage est prise comme référence ($f_e = 1$) et dont la réponse en fréquence $H(f)$ est telle que :

$$H(k\,.\,0{,}0625) = 1 \quad \text{pour} \quad k = 0, 1, 2, 3.$$
$$H(0{,}25) = 0{,}5$$
$$H(k\,.\,0{,}0625) = 0 \quad \text{pour} \quad k = 5, 6, 7, 8.$$

Calculer par transformation de Fourier Discrète les 17 coefficients de ce filtre. Tracer la réponse en fréquence et donner les zéros de la fonction de transfert en Z.

3 Utiliser les formules du paragraphe 5.7 pour déterminer les ondulations d'un filtre passe-bas à 17 coefficients dont la fréquence de fin de bande passante est donnée par $f_1 = 0{,}2$, et la fréquence de début de bande affaiblie par $f_2 = 0{,}3$. Comparer les résultats obtenus à ceux des exercices précédents.

4 Soit un filtre dont la fonction de transfert $H(f)$ est donnée, à un déphasage près, par l'équation :

$$H(f) = h_0 + 2 \sum_{i=1}^{4} h_{2i-1} \cos\left[2\pi f (2i-1)\, T\right]$$

Donner les structures directe et transposée permettant de réaliser ce filtre avec le minimum d'éléments. Quelles simplifications interviennent si la fréquence d'échantillonnage de sortie peut être divisée par deux ?

5 Un filtre passe-bas étroit est défini par l'équation :

$$y(n) = \sum_{i=0}^{N-1} a_i x(n-i)$$

Comment se trouve modifiée la réponse en fréquence si les coefficients a_i sont remplacés par : $a_i(-1)^i$ et par $a_i \cos\left(\frac{i\pi}{2}\right)$? Que deviennent les zéros du filtre dans cette opération.

6 Soit un filtre passe-bas satisfaisant au gabarit de la figure 5.7 avec les valeurs de paramètres :

$$f_1 = 0{,}05 ; \quad f_2 = 0{,}15 ; \quad \delta_1 = 0{,}01 \quad \text{et} \quad \delta_2 = 0{,}001.$$

Combien de coefficients sont nécessaires et combien de bits faut-il pour les représenter ? Si les données appliquées à ce filtre ont 12 bits, si la dégradation tolérable du rapport signal à bruit est limitée à $\Delta SB = 0{,}1$ dB, combien de bits doivent avoir les données internes ?

7 Donner l'expression de la réponse en fréquence du filtre de l'exemple du paragraphe 5.3. Vérifier la réponse en fréquence aux points 0, 0,25 et 0,5.

Les coefficients sont arrondis à 6 bits (signe compris). Donner l'expression de la fonction erreur $e(f)$ introduite et la calculer au voisinage du point $f = 0{,}1925$ de l'axe des fréquences. Élaborer une formule analogue à (5.46) pour l'estimation du nombre de bits nécessaire pour représenter les coefficients dans ce type de filtre, en suivant la démarche du paragraphe 5.10.

Chapitre **6**

Cellules de filtres à réponse impulsionnelle infinie (RII)

Les filtres numériques à réponse impulsionnelle infinie sont des systèmes linéaires discrets invariants dans le temps dont le fonctionnement est régi par une équation de convolution portant sur une infinité de termes. En principe, ils conservent une trace des signaux qui leur ont été appliqués pendant une durée infinie, ils sont à mémoire infinie. Une telle mémoire est réalisée par une boucle de réaction de la sortie sur l'entrée, d'où la dénomination courante de filtre récursif. Chaque élément de la suite des nombres de sortie est calculé par sommation pondérée d'un certain nombre d'éléments de la suite d'entrée et d'un certain nombre d'éléments de la suite de sortie précédents.

Le fait d'avoir cette réponse impulsionnelle infinie permet d'obtenir en général des fonctions de filtrage beaucoup plus sélectives que celles des filtres RIF à quantité de calculs équivalente. Cependant la boucle de réaction complique l'étude des propriétés et la conception de ces filtres et amène des phénomènes parasites.

Pour aborder l'étude des filtres RII, il est plus simple de considérer d'abord les cellules de filtres élémentaires du premier et du second ordre. En fait, l'intérêt de ces structures simples va bien au-delà d'une introduction aux propriétés des filtres RII, car elles constituent la forme de réalisation la plus courante. En effet, c'est en général sous la forme d'un ensemble de telles cellules élémentaires que se présentent en pratique les filtres RII, même les plus complexes.

6.1 LA CELLULE ÉLÉMENTAIRE DU PREMIER ORDRE

Soit le système qui, à la suite de données $x(n)$, fait correspondre la suite $y(n)$ telle que :

$$y(n) = x(n) + by(n-1) \tag{6.1}$$

où b est une constante.

C'est une cellule élémentaire du premier ordre.

La réponse de ce système à la suite unitaire $u_0(n)$ telle que :

$$\begin{aligned} u_0(n) &= 1 \quad \text{pour} \quad n = 0 \\ u_0(n) &= 0 \quad \text{pour} \quad n \neq 0 \end{aligned}$$

est la suite $y_0(n)$ telle que :

$$\begin{aligned} y_0(n) &= 0 \quad \text{pour} \quad n < 0 \\ y_0(n) &= b^n \quad \text{pour} \quad n \geqslant 0 \end{aligned}$$

Cette suite constitue la réponse impulsionnelle du filtre, elle est définie et la condition de stabilité s'écrit :

$$\sum_{n=0}^{\infty} |b|^n < \infty$$

d'où : $|b| < 1$.

La réponse du système à la suite $x(n)$ telle que :

$$\begin{aligned} x(n) &= 0 \quad \text{pour} \quad n < 0 \\ x(n) &= 1 \quad \text{pour} \quad n \geqslant 0 \end{aligned}$$

est la suite $y(n)$ telle que :

$$\begin{aligned} y(n) &= 0 \quad && \text{pour} \quad n < 0 \\ y(n) &= \frac{1 - b^{n+1}}{1 - b} \quad && \text{pour} \quad n \geqslant 0 \end{aligned} \tag{6.2}$$

qui tend vers $\dfrac{1}{1-b}$ quand n tend vers l'infini, si le système est stable.

Cette réponse est représentée sur la figure 6.1.

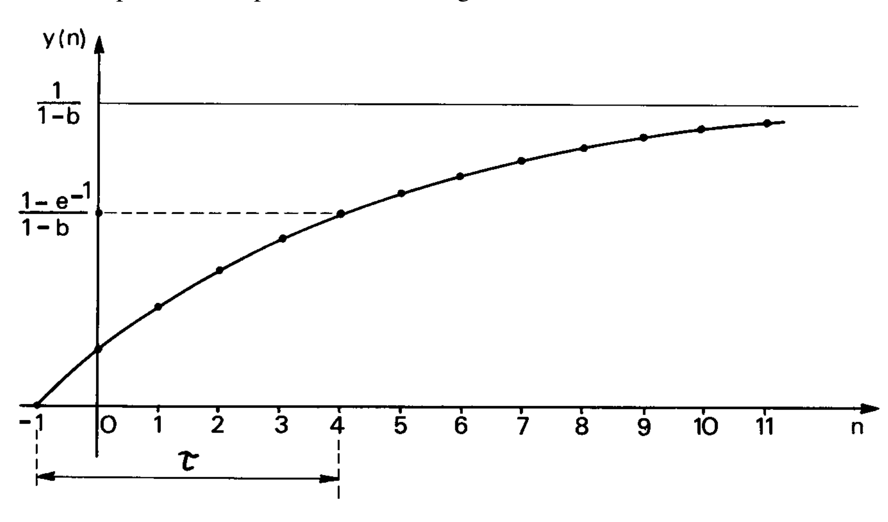


FIG. 6.1. *Réponse de la cellule du premier ordre à l'échelon unité*

Par analogie avec le système continu de constante de temps τ, échantillonné avec la période T et dont la réponse $y_c(n)$ s'écrit :

$$y_c(n) = \left[1 - e^{-\frac{T}{\tau}(n+1)}\right]$$

on définit la constante de temps de la cellule numérique du premier ordre τ en égalant les valeurs $y_c(0) = (1-b)\,y(0)$. D'où :

$$y_c(0) = (1-b)\,y(0) = 1-b \quad \text{soit :} \quad e^{-\frac{T}{\tau}} = b$$

pour $b > 0$. Il vient alors :

$$\tau = \frac{T}{\ln\left(\frac{1}{b}\right)} \tag{6.3}$$

Pour des valeurs de b proches de l'unité :

$$b = 1 - \delta \quad \text{avec} \quad 0 < \delta \ll 1$$

c'est-à-dire les systèmes définis par la relation :

$$y(n) = x(n) + (1-\delta)\,y(n-1) \tag{6.4}$$

il vient :

$$\tau \simeq \frac{T}{\delta} \tag{6.5}$$

Cette situation se rencontre dans les systèmes adaptatifs étudiés au chapitre 14.

Si la suite $x(n)$ résulte, pour $n \geqslant 0$, de l'échantillonnage du signal $x(t) = e^{j2\pi ft}$ ou $x(t) = e^{j\omega t}$, avec la période T = 1, il vient :

$$y(n) = \frac{e^{jn\omega}}{1 - be^{-j\omega}} - \frac{b^{n+1}e^{-j\omega}}{1 - be^{-j\omega}} \tag{6.6}$$

Cette expression fait apparaître un régime transitoire et un régime permanent qui correspond à la réponse en fréquence $H(\omega)$ du filtre :

$$H(\omega) = \frac{1}{1 - be^{-j\omega}} \tag{6.7}$$

En faisant apparaître le module et la phase de cette fonction il vient :

$$|H(\omega)|^2 = \frac{1}{1 - 2b\cos\omega + b^2} \; ; \quad \varphi(\omega) = \text{Arctg}\,\frac{b\sin\omega}{1 - b\cos\omega} \tag{6.8}$$

Pour le temps de propagation de groupe :

$$\tau_g(\omega) = \frac{d\varphi}{d\omega} = \frac{b\cos\omega - b^2}{1 - 2b\cos\omega + b^2} \tag{6.9}$$

On peut remarquer que pour ω très petit il est possible d'écrire :

$$|H(\omega)|^2 \simeq \frac{1}{(1-b)^2\left[1+\dfrac{b}{(1-b)^2}\,\omega^2\right]} \tag{6.10}$$

Cette expression est à rapprocher de la réponse $H_{RC}(\omega)$ d'un circuit RC qui s'écrit :

$$|H_{RC}(\omega)|^2 = \frac{1}{1+R^2C^2\omega^2} \qquad (6.10\text{-}bis)$$

Il apparaît que pour les fréquences très faibles devant la fréquence d'échantillonnage, le circuit numérique a une réponse qui peut être assimilée à celle d'un réseau RC.

La figure 6.2.*a* représente la forme de la réponse en fréquence du circuit numérique du premier ordre. La figure 6.2.*b* donne la réponse en phase et la figure 6.2.*c* le temps de groupe.

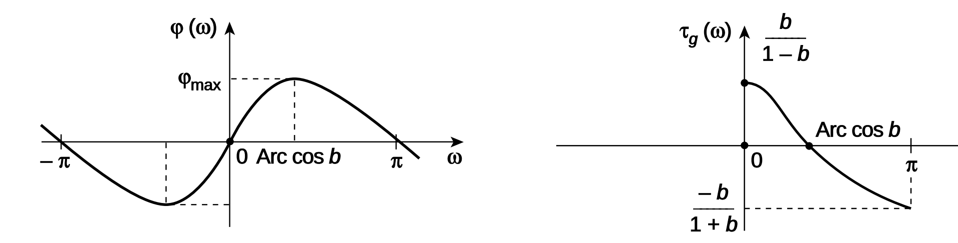


FIG. 6.2.*b*. *Réponse en phase*

FIG. 6.2.*c*. *Temps de groupe*

La phase peut encore s'écrire :

$$\varphi(\omega) = \text{Arc tg}\,\frac{\sin\omega}{\cos\omega - b} - \omega; \quad \cos\omega > b$$

$$\varphi_{max} = \frac{\pi}{2} - \text{Arc cos}\, b; \quad \cos\omega = b \tag{6.11}$$

$$\varphi(\omega) = \pi + \text{Arc tg}\,\frac{\sin\omega}{\cos\omega - b} - \omega; \quad \cos\omega < b$$

Elle passe donc par un maximum pour ω tel que cos ω = *b*, ce qui correspond à l'annulation du temps de groupe. Le coefficient *b* contrôle donc ainsi directement le maximum de la phase de la cellule.

La fonction de transfert de la cellule du 1er ordre s'obtient aussi à l'aide de la transformée en Z. Soit Y(Z) et X(Z) les transformées des suites de sortie et d'entrée respectivement; il vient :

$$Y(Z) = X(Z) + b\,Z^{-1}\,Y(Z)$$

d'où la fonction de transfert en Z, H(Z) telle que :

$$H(Z) = \frac{1}{1 - bZ^{-1}} = \frac{Z}{Z - b}$$

La réponse en fréquence s'obtient simplement en remplaçant Z par $e^{j\omega}$, dans l'expression de H(Z), avec $\omega = 2\pi f$.

L'interprétation graphique conduit à la figure 6.2.*d.* qui représente le pôle P de cette fonction dans le plan complexe ; c'est un point de l'axe réel, d'abscisse *b*.

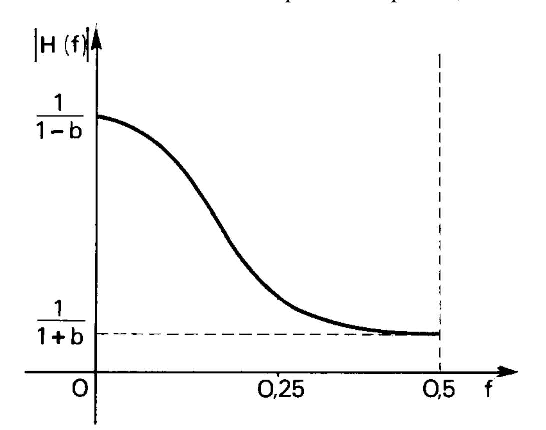


FIG. 6.2.*a.* *Réponse en fréquence de la cellule du premier ordre*

FIG. 6.2.*d.* *Pôle de la cellule du premier ordre*

Conformément à cette figure :

$$|H| = \frac{1}{MP} \quad \text{et} \quad \varphi = \alpha - \omega$$

La condition de stabilité implique que le pôle P soit à l'intérieur du cercle unité.

Un cas particulier intéressant est celui de l'intégrateur à bande étroite, défini par la fonction de transfert suivante :

$$H_{int}(Z) = \frac{\varepsilon}{1 - (1 - \varepsilon)\, Z^{-1}} \tag{6.12}$$

avec ε petit tel que $0 < \varepsilon \ll 1$.

On peut montrer alors que la largeur de bande à 3 dB est approximativement égale à ε et la constante de temps égale à $1/\varepsilon$. Quant à la norme de la réponse en fréquence, elle s'écrit : $\|H\|_2^2 \simeq \frac{\varepsilon}{2}$.

La transformée en Z monolatérale permet de faire apparaître les régimes transitoires et d'introduire les conditions initiales En effet :

$$\sum_{n=0}^{\infty} y(n) Z^{-n} = \sum_{n=0}^{\infty} x(n) Z^{-n} + b \sum_{n=0}^{\infty} y(n-1) Z^{-n}$$

$$Y(Z) = X(Z) + by(-1) + bZ^{-1} Y(Z)$$

d'où

$$Y(Z) = \frac{X(Z)}{1 - bZ^{-1}} + \frac{by(-1)}{1 - bZ^{-1}}$$

si $x(n) = e^{jn\omega}$, X(Z) s'écrit :

$$X(Z) = \sum_{n=0}^{\infty} e^{jn\omega} Z^{-n} = \frac{1}{1 - e^{j\omega} Z^{-1}} \qquad (6.13)$$

La valeur $y(n)$ s'obtient par la formule de la transformée en Z inverse :

$$y(n) = \frac{1}{j2\pi} \int_{\Gamma} Z^{n-1} \left[\frac{1}{1 - e^{j\omega} Z^{-1}} \cdot \frac{1}{1 - bz^{-1}} + \frac{by(-1)}{1 - bZ^{-1}} \right] dZ$$

En prenant comme contour d'intégration Γ un cercle de rayon supérieur à l'unité, le théorème des résidus donne :

$$y(n) = \frac{e^{jn\omega}}{1 - be^{-j\omega}} - \frac{b^{n+1} e^{-j\omega}}{1 - be^{-j\omega}} + y(-1)\, b^{n+1} \qquad (6.14)$$

Cette expression peut aussi être obtenue de manière directe par développement en série de Y(Z). Elle fait apparaître, en plus de la réponse correspondant au régime permanent, la réponse transitoire et la réponse due aux conditions initiales. Ces dernières disparaissent quand n croît si $|b| < 1$, c'est-à-dire si le système est stable.

De cette analyse, il résulte que la cellule du premier ordre offre des possibilités restreintes car elle ne possède qu'un pôle, qui doit être réel pour que le filtre soit à coefficients réels, et sa réponse en fréquence est une fonction monotone. La cellule du second ordre offre des possibilités beaucoup plus variées. C'est la structure la plus utilisée en filtrage numérique en raison de la modularité qu'elle apporte dans la réalisation des filtres même les plus complexes et de ses propriétés concernant la limitation du nombre de bits des coefficients et le bruit de calcul.

Le cas de la cellule qui ne comporte que des pôles, ou cellule purement récursive, va être examiné d'abord.

6.2 LA CELLULE DU SECOND ORDRE PUREMENT RÉCURSIVE

Soit un système qui à la suite de données $x(n)$ fait correspondre la suite $y(n)$ telle que :

$$y(n) = x(n) - b_1 y(n-1) - b_2 y(n-2) \qquad (6.15)$$

Dans cette expression, le signe des coefficients b_1 et b_2 est changé par rapport au paragraphe précédent, pour faciliter l'écriture de la fonction de transfert en Z du système, H(Z), donnée par :

$$H(Z) = \frac{1}{1 + b_1 Z^{-1} + b_2 Z^{-2}} = \frac{Z^2}{Z^2 + b_1 Z + b_2}$$

Cette fonction possède un zéro double à l'origine et deux pôles P_1 et P_2, tels que :

$$P_{1,2} = -\frac{b_1}{2} \pm \frac{1}{2}\sqrt{b_1^2 - 4b_2} \qquad (6.16)$$

Deux cas se présentent alors suivant le signe de $b_1^2 - 4b_2$:

• $b_1^2 \geqslant 4b_2$: les deux pôles sont situés sur l'axe réel du plan complexe ; la fonction de transfert est simplement le produit de deux fonctions du premier ordre à coefficients réels. La cellule de filtre correspondante est la mise en cascade de deux cellules du premier ordre et ses propriétés s'en déduisent.

Les amplitudes se multiplient et les phases s'ajoutent. La réponse à l'échelon unité en sortie de la seconde cellule s'écrit, si b_1 et b_2 désignent les coefficients :

$$y_2(n) = \frac{1}{(1-b_1)(1-b_2)}\left[1 - b_2^{n+1} - (1-b_2)\frac{b_1^{n+1} - b_2^{n+1}}{b_1 - b_2}\right] \qquad (6.17)$$

La constante de temps correspondante τ_{12} s'écrit, pour des coefficients proches de l'unité :

$$\tau_{12} \simeq \sqrt{2}\sqrt{\tau_1\tau_2} \qquad (6.18)$$

et pour des cellules identiques :

$$\tau_{12} \simeq \sqrt{2}\,\tau_1$$

Plus généralement, pour N cellules identiques, la constante de temps τ_N s'exprime approximativement par :

$$\tau_N \simeq \sqrt{N}\,\tau_1 \qquad (6.19)$$

• $b_1^2 < 4b_2$: les deux pôles sont complexes conjugués ; ils s'écrivent P et $\overline{P}$, avec :

$$P = -\frac{b_1}{2} + j\,\frac{1}{2}\sqrt{4b_2 - b_1^2} \qquad (6.20)$$

La figure 6.3 illustre ce cas qui est le plus intéressant, et est considéré exclusivement dans la suite de ce paragraphe.

La relation entre la position des pôles et les coefficients du filtre apparaît très simplement :

$$b_1 = -2\mathrm{Re}(P) \qquad (6.21)$$

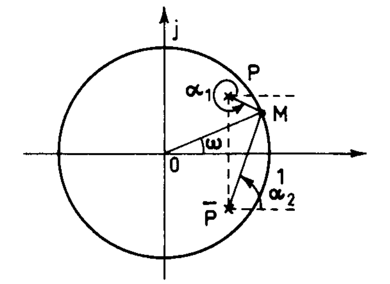


FIG. 6.3. *Cellule du second ordre à pôles complexes*

c'est-à-dire que le coefficient du terme en Z^{-1} dans l'expression de $H(Z)$ est égal en module à deux fois la partie réelle du pôle et de signe inverse.

$$b_2 = |OP|^2 \tag{6.22}$$

Le coefficient du terme en Z^{-2} est égal au carré du module du pôle ou encore au carré de la distance du pôle à l'origine. Ces deux relations sont très utiles dans la détermination des coefficients des filtres comme on le verra dans la suite.

Si M désigne le point d'affixe $e^{j\omega}$ dans le plan complexe, le module de la fonction de transfert s'écrit :

$$|H(\omega)| = \frac{1}{MP \cdot M\bar{P}}$$

et la phase :

$$\varphi(\omega) = \alpha_1 + \alpha_2 - 2\omega$$

où α_1 et α_2 désignent les angles que font les vecteurs $\overrightarrow{PM}$ et $\overrightarrow{\bar{P}M}$ avec l'axe réel.

Les expressions analytiques se déduisent de $H(Z)$ en faisant $Z = e^{j\omega}$. En prenant pour $H(Z)$ l'expression :

$$H(Z) = \frac{1}{1 + b_1 Z^{-1} + b_2 Z^{-2}}$$

il vient :

$$|H(\omega)|^2 = \frac{1}{1 + b_1^2 + b_2^2 + 2b_1(1 + b_2)\cos\omega + 2b_2\cos 2\omega} \tag{6.23}$$

$$\varphi(\omega) = -\,\mathrm{Arctg}\left[\frac{b_1 \sin\omega + b_2 \sin 2\omega}{1 + b_1 \cos\omega + b_2 \cos 2\omega}\right] \tag{6.24}$$

Une forme élégante pour exprimer la réponse en fréquence et la phase est obtenue avec une représentation des pôles en coordonnées polaires, $P = re^{j\theta}$, à partir d'une expression de $H(Z)$ en produit de facteurs :

$$H(Z) = \frac{1}{(1 - PZ^{-1})(1 - \bar{P}Z^{-1})}$$

Les relations avec les coefficients b_1 et b_2 sont les suivantes :

$$b_1 = -2r\cos\theta\,; \quad b_2 = r^2$$

Pour $H(\omega)$ on obtient :

$$H(\omega) = \frac{1}{[1 - re^{j(\theta-\omega)}][1 - re^{-j(\theta-\omega)}]} \tag{6.25}$$

Il vient alors :

$$|H(\omega)|^2 = \frac{1}{[1 + r^2 - 2r\cos(\theta - \omega)][1 + r^2 - 2r\cos(\theta + \omega)]} \tag{6.23-\textit{bis}}$$

$$\varphi(\omega) = \mathrm{Arctg}\left[\frac{r\sin(\theta+\omega)}{1 - r\cos(\theta+\omega)}\right] - \mathrm{Arctg}\left[\frac{r\sin(\theta-\omega)}{1 - r\cos(\theta-\omega)}\right] \tag{6.24-\textit{bis}}$$

Ces expressions permettent de tracer les courbes donnant $|H(\omega)|$ et $\varphi(\omega)$ en fonction de la pulsation $\omega = 2\pi f$.

On vérifie que $|H(\omega)|$ est une fonction paire et que $\varphi(\omega)$ est une fonction impaire de la variable ω.

Les valeurs correspondant à des extremums de $|H(\omega)|$ sont les racines de l'équation suivante, obtenue en dérivant l'expression (6.23) par rapport à ω :

$$\sin \omega [b_1(1 + b_2) + 4b_2 \cos \omega] = 0$$

Les fréquences 0 et 0,5 sont des fréquences extrémales en raison de la symétrie et de la périodicité de la réponse. Une autre fréquence extrémale f_0 existe si la condition suivante est remplie :

$$\left| \frac{b_1(1 + b_2)}{4b_2} \right| < 1 \qquad (6.26)$$

ou encore en coordonnées polaires :

$$\cos \theta < \frac{2r}{1 + r^2} \qquad (6.26\text{-}bis)$$

Dans ce cas il vient :

$$\cos (2\pi f_0) = \cos \omega_0 = -\frac{b_1(1 + b_2)}{4b_2} \qquad (6.27)$$

La fréquence f_0 est la fréquence de résonance de la cellule. L'amplitude à la résonance s'écrit :

$$H_m = \frac{1}{1 - b_2} \sqrt{\frac{4b_2}{4b_2 - b_1^2}} \qquad (6.28)$$

ou encore en coordonnées polaires :

$$H_m = \frac{1}{1 - r} \cdot \frac{1}{(1 + r) \sin \theta} \qquad (6.29)$$

Il apparaît ainsi que la réponse en fréquence à la résonance est inversement proportionnelle à la distance du pôle au cercle unité. Cette expression constitue un résultat fondamental, souvent utilisé par la suite.

Il est intéressant également de faire apparaître pour la cellule du second ordre la caractéristique appelée largeur de bande à 3 décibels, B_3, telle que :

$$B_3 = f_2 - f_1 = \frac{\omega_2 - \omega_1}{2\pi}$$

avec : $$|H(\omega_1)|^2 = |H(\omega_2)|^2 = \frac{1}{2} H_m^2$$

Pour une cellule à forte résonance ($r \simeq 1$), d'après (6.22) et (6.23) on peut écrire approximativement au voisinage de la fréquence de résonance :

$$|H(\omega_1)|^2 \simeq \frac{1}{4 \sin^2 \theta} \cdot \frac{1}{1 + r^2 - 2r \cos (\theta - \omega_1)} = \frac{1}{2} \frac{1}{2(1 - r^2) \sin^2 \theta}$$

d'où :
$$\cos(\theta - \omega_1) = \frac{1 + r^2}{2r} - \frac{(1 - r^2)^2}{4r}$$

Par développement limité, on obtient :

$$|\theta - \omega_1| \simeq 1 - r$$

D'où l'approximation pour une cellule à forte résonance :

$$B_3 = \frac{1 - r}{\pi} \qquad (6.30)$$

Ce résultat est utilisé par la suite dans les calculs de complexité.

Une autre caractéristique est parfois utilisée pour une cellule du second purement récursive, la bande équivalente du bruit B_2. C'est la largeur de bande d'un bruit dont la densité spectrale est supposée constante dans cette bande et égale à H_m^2 et dont la puissance totale est égale à la puissance obtenue en sortie de la cellule quand un bruit blanc de puissance unitaire est appliqué. Par définition :

$$B_b \,.\, H_m^2 = \|H\|_2^2$$

En tenant compte de l'expression de $\|H\|_2^2$ donnée ci-dessous (6.36) et de la relation (6.29) ci-dessus, il vient :

$$B_b = \frac{(1 - r^2)\sin^2\theta}{1 + r^4 - 2r^2\cos 2\theta} \qquad (6.30\text{-}bis)$$

Cette expression est utile en analyse spectrale par exemple.

Les caractéristiques principales de la cellule du second ordre purement récursive sont illustrées par un exemple.

Exemple

Soit une cellule du second ordre dont les pôles ont pour affixe :

$$P = 0{,}6073 + j\,0{,}5355$$
$$\overline{P} = 0{,}6073 - j\,0{,}5355$$

Les paramètres sont les suivants :

$$b_1 = -\,2Re\,(P) = -\,1{,}2146$$
$$b_2 = |OP|^2 = 0{,}6556$$
$$H(Z) = \frac{1}{1 - 1{,}2146\,Z^{-1} + 0{,}6556\,Z^{-2}}$$
$$|H(\omega)|^2 = \frac{1}{2{,}905 - 4{,}02\cos\omega + 1{,}31\cos(2\omega)}$$
$$\theta = 2\pi\,.\,0{,}1156\,; \quad r = 0{,}81\,; \quad f_0 = 0{,}111\,; \quad H_m = 4{,}39\,; \quad B_3 = 0{,}06$$

Le module de la réponse est représenté sur la figure 6.4 en fonction de la fréquence.

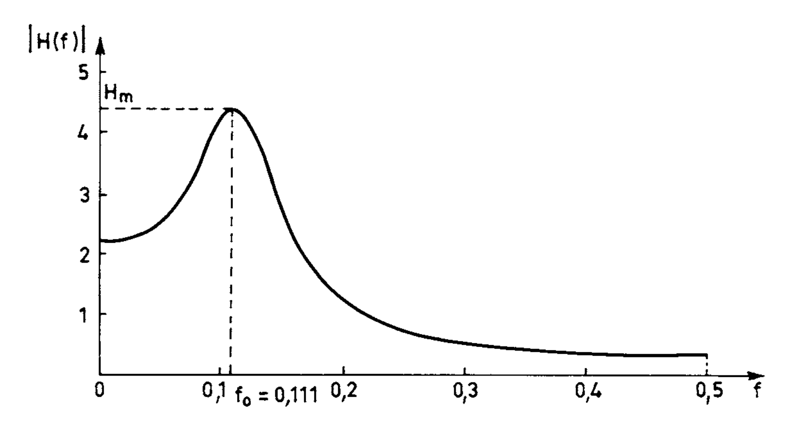


FIG. 6.4. *Réponse d'une cellule du 2e ordre purement récursive*

La réponse en phase de la cellule du second ordre s'étudie à partir des relations (6.24) qui expriment la fonction $\varphi(\omega)$. Pour préciser les variations de cette fonction il est utile de calculer d'abord sa dérivée, c'est-à-dire le temps de groupe. A partir de la relation (6.24-*bis*), on obtient :

$$\frac{d\varphi}{d\omega} = \frac{r\cos(\theta+\omega) - r^2}{1 - 2r\cos(\theta+\omega) + r^2} + \frac{r\cos(\theta-\omega) - r^2}{1 - 2r\cos(\theta-\omega) + r^2} \tag{6.31}$$

Soit :

$$\tau(\omega) = \frac{r\,[\cos(\theta+\omega) - r]}{1 - 2r\cos(\theta+\omega) + r^2} + \frac{r\,[\cos(\theta-\omega) - r]}{1 - 2r\cos(\theta-\omega) + r^2} \tag{6.32}$$

La fonction $\tau(\omega)$ passe par un maximum au voisinage de la fréquence de résonance. A la fréquence $f = \dfrac{\theta}{2\pi}$ il vient :

$$\tau(\theta) = \frac{r}{1-r}\left[1 + \frac{(1-r)\,[\cos 2\theta - r]}{1 - 2r\cos 2\theta + r^2}\right] \simeq \frac{r}{1-r} \tag{6.33}$$

Exemple

$$r = 0{,}81\,; \quad \theta = 2\pi\,.\,0{,}1156$$

La figure 6.5 donne la courbe $\tau(f)$ en fonction de la fréquence. Cette courbe passe par un maximum égal à 3,8 au voisinage de la résonance. L'unité de temps est la période d'échantillonnage T. Les valeurs obtenues sont à multiplier par T si cette période est différente de l'unité.

Il apparaît que la fonction $\tau(f)$ prend des valeurs négatives. En fait il s'agit du temps de propagation de groupe théorique de la cellule. En effet chaque élément de sortie $y(n)$ est calculé par une addition où intervient un nombre d'entrée $x(n)$ et cette opération ne peut être instantanée. Pour rendre le système réalisable il faut

retarder $y(n)$, par exemple d'une unité; le temps de groupe se trouve alors augmenté d'autant: à la phase $\varphi(\omega)$, il faut ajouter la valeur ω. La fonction $\varphi(\omega)$ obtenue dans ces conditions est représentée sur la figure 6.6; la pente est maximale au voisinage de la résonance.

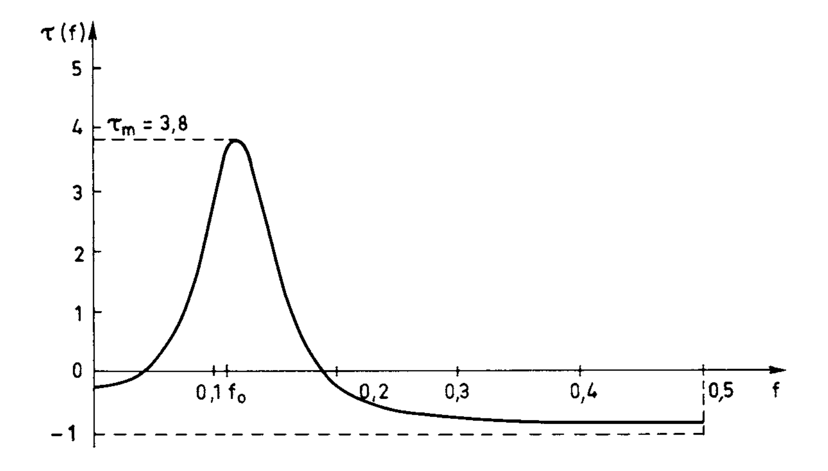


Fig. 6.5. *Temps de propagation de groupe théorique de la cellule purement récursive*

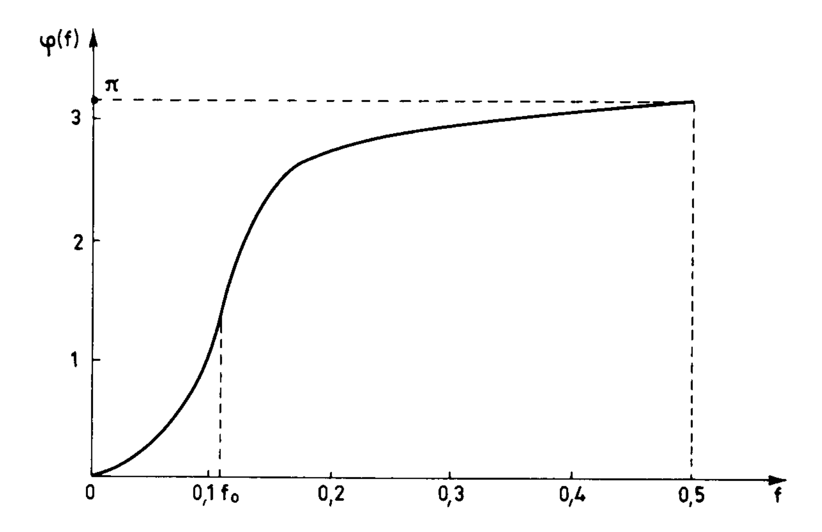


Fig. 6.6. *Caractéristique de phase de la cellule purement récursive*

Les expressions qui ont été données pour les fonctions $|H(\omega)|$, $\varphi(\omega)$ et $\tau(\omega)$ sont importantes car elles permettent de déterminer les mêmes fonctions pour les filtres réalisés par mise en cascade de cellules du second ordre, par multiplication pour le module de la réponse en fréquence ou par addition pour la phase et le temps de propagation de groupe.

Pour faire apparaître les conditions initiales et les régimes transitoires, la transformation en Z monolatérale est utilisée, comme précédemment. A partir de l'équation de définition de la cellule, on obtient la relation suivante entre les transformées monolatérales Y (Z) et X (Z) :

$$Y(Z) = \frac{X(Z)}{1 + b_1 Z^{-1} + b_2 Z^{-2}} - \frac{b_1 y(-1) + b_2 [y(-2) + y(-1) Z^{-1}]}{1 + b_1 Z^{-1} + b_2 Z^{-2}} \tag{6.34}$$

Pour $x(n) = e^{jn\omega}$, on obtient $y(n)$ par la formule :

$$y(n) = \frac{1}{j2\pi} \int_{\Gamma} Z^{n-1} Y(Z)\, dZ$$

avec

$$X(Z) = \frac{1}{1 - e^{j\omega} Z^{-1}}$$

en prenant comme contour d'intégration Γ un cercle de rayon supérieur à l'unité.

L'étude de la cellule purement récursive a été faite dans le plan fréquentiel. Dans le plan temporel cette cellule possède une réponse impulsionnelle qui est une suite $h(n)$ que l'on détermine directement en examinant la réponse à la suite unitaire, ou par développement en série de la fonction H (Z) ; en effet on a, pour des pôles complexes :

$$H(Z) = \frac{1}{1 - PZ^{-1}} \cdot \frac{P}{P - \overline{P}} + \frac{1}{1 - \overline{P}Z^{-1}} \cdot \frac{\overline{P}}{P - \overline{P}} = \sum_{n=0}^{\infty} h(n) Z^{-n}$$

Il vient alors :

$$h(n) = r^n \cdot \frac{\sin(n+1)\theta}{\sin\theta} \tag{6.35}$$

La figure 6.7 donne la réponse impulsionnelle du filtre de l'exemple précédent.

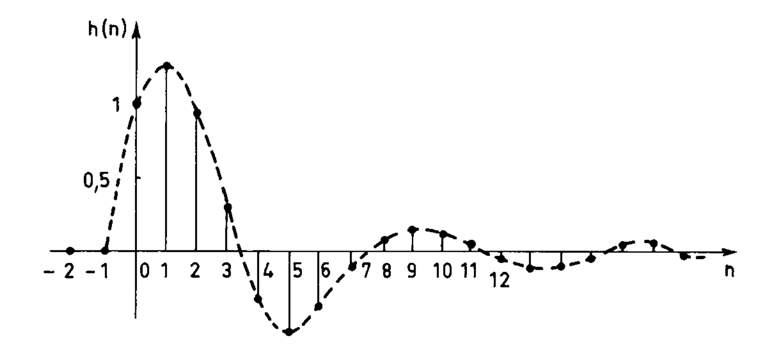


FIG. 6.7. *Réponse impulsionnelle d'une cellule du second ordre*

La réponse à l'échelon unité s'écrit à partir de la définition, après quelques manipulations :

$$g(n) = \frac{1}{1 + b_1 + b_2}\,[1 + b_2 h(n) - h(n+1)]$$

Il vient alors :

$$g(n) = \frac{1}{1 + r^2 - 2r\cos\theta}\left[1 + \frac{r^{i+1}}{\sin\theta}\,[r\sin(i+1)\theta - \sin(i+2)\theta]\right] \qquad (6.34\text{-}bis)$$

Cette relation est utile en automatique.

La norme $\|H\|_2$ de la fonction $H(\omega)$ est utilisée par la suite ; le calcul de cette norme peut se faire par deux méthodes, comme indiqué au paragraphe 4.3. Par sommation de série :

$$\|H\|_2^2 = \sum_{n=0}^{\infty} |h(n)|^2 = \frac{1}{\sin^2\theta}\sum_{n=0}^{\infty} r^{2n}\,\frac{1 - \cos[2(n+1)\theta]}{2}$$

Par un calcul d'intégrale suivant la méthode des résidus :

$$\|H\|_2^2 = \frac{1}{j2\pi}\int_{|Z|=1} \frac{Z\,dZ}{(Z - P)(Z - \bar{P})(1 - PZ)(1 - \bar{P}Z)}$$

Finalement il vient :

$$\|H\|_2^2 = \frac{1 + r^2}{1 - r^2}\cdot\frac{1}{1 + r^4 - 2r^2\cos 2\theta} \qquad (6.36)$$

La valeur $\|H\|_1$ est également utilisée par la suite :

$$\|H\|_1 = \sum_{n=0}^{\infty} |h(n)|$$

Cette valeur est bornée par l'inégalité :

$$\|H\|_1 = \frac{1}{\sin\theta}\sum_{n=0}^{\infty} r^n|\sin[(n+1)\theta]| \leqslant \frac{1}{(1-r)\sin\theta} \quad (0 < \theta < \pi) \qquad (6.37)$$

Exemple :

Quand les pôles sont sur l'axe imaginaire du plan des Z, $\theta = \frac{\pi}{2}$ et la réponse impulsionnelle s'écrit :

$$h(2p) = r^{2p}(-1)^p$$

Alors :

$$\|H\|_2^2 = \sum_{p=0}^{\infty} |h(2p)|^2 = \frac{1}{1 - r^4}$$

$$\|H\|_1 = \sum_{p=0}^{\infty} |h(2p)| = \frac{1}{1 - r^2}$$

Les résultats obtenus pour la cellule du second ordre purement récursive s'étendent à la cellule du second ordre générale.

6.3 CELLULE DU SECOND ORDRE GÉNÉRALE

La cellule du second ordre la plus générale fait intervenir dans le calcul d'un élément de la suite de sortie $y(n)$ à l'instant n, les données aux instants précédents, $x(n-1)$ et $x(n-2)$. Son équation de définition s'écrit :

$$y(n) = a_0 x(n) + a_1 x(n-1) + a_2 x(n-2) - b_1 y(n-1) - b_2 y(n-2) \qquad (6.38)$$

Il en résulte la fonction de transfert en Z suivante :

$$H_T(Z) = \frac{a_0 + a_1 Z^{-1} + a_2 Z^{-2}}{1 + b_1 Z^{-1} + b_2 Z^{-2}}$$

qui comporte deux zéros réels ou complexes conjugués pour que les coefficients soient réels. La position de ces zéros, notés Z_0 et $\overline{Z_0}$, est assez particulière. En effet on rencontre deux cas d'utilisation de la cellule du second ordre générale. D'abord dans la réalisation d'un élément de filtrage ; alors les zéros sont presque toujours placés sur le cercle unité, d'une part pour optimiser les caractéristiques d'affaiblissement du filtre par l'introduction d'une fréquence d'affaiblissement infini et d'autre part parce que dans ces conditions une symétrie des coefficients apparaît et les calculs peuvent se simplifier. Ensuite dans la réalisation de circuits déphaseurs purs ; alors les zéros sont conjugués harmoniques des pôles.

Le cas du filtre va être examiné en premier. La fonction de transfert d'une cellule s'écrit :

$$H_T(z) = \frac{a_0(1 + a_1 z^{-1} + z^{-2})}{1 + b_1 z^{-1} + b_2 z^{-2}} = a_0 \frac{(Z - Z_0)(Z - \overline{Z_0})}{(Z - P)(Z - \overline{P})} \qquad (6.39)$$

ou encore :

$$H_T(Z) = a_0 \frac{1 - 2Re(Z_0) Z^{-1} + Z^{-2}}{1 - 2Re(P) Z^{-1} + |P|^2 Z^{-2}}$$

Le module de la réponse en fréquence de la cellule du second ordre générale dont les zéros sont placés sur le cercle unité s'exprime par :

$$|H_T(\omega)|^2 = \frac{(a_1 + 2a_0 \cos\omega)^2}{1 + b_1^2 + b_2^2 + 2b_1(1 + b_2)\cos\omega + 2b_2 \cos 2\omega} \qquad (6.40)$$

Une telle cellule peut être considérée comme la mise en cascade d'une cellule de filtre RII purement récursive et d'une cellule de filtre RIF à phase linéaire. Par suite les caractéristiques de phase et de temps de groupe de la cellule complète sont les sommes des caractéristiques des cellules élémentaires. C'est-à-dire :

$$\tau_T(\omega) = 1 + \frac{r\cos(\theta + \omega) - r^2}{1 - 2r\cos(\theta + \omega) + r^2} + \frac{r\cos(\theta - \omega) - r^2}{1 - 2r\cos(\theta - \omega) + r^2} \qquad (6.41)$$

$$\varphi_T(\omega) = \omega + \text{Arctg}\left[\frac{r \sin(\theta + \omega)}{1 - r\cos(\theta + \omega)}\right] - \text{Arctg}\left[\frac{r\sin(\theta - \omega)}{1 - r\sin(\theta - \omega)}\right] \tag{6.42}$$

Ces deux expressions donnent la phase et le temps de propagation de groupe d'une cellule du second ordre dont les deux zéros sont sur le cercle unité. Cette cellule de filtrage est généralement appelée cellule du second ordre elliptique, par référence à la technique utilisée pour le calcul des coefficients.

Exemple

Pour illustrer les propriétés de la cellule de filtrage du second ordre générale, reprenons l'exemple du paragraphe précédent, en complétant le filtre par 2 zéros tels que :

$$Z_0 = 0{,}3325 + j\,0{,}943 \quad \text{et} \quad \overline{Z_0} = 0{,}3325 - j\,0{,}943$$

Les positions des singularités dans le plan complexe sont données par la figure 6.8.a. La fonction de transfert $H_T(Z)$ est le quotient de deux polynômes du second degré $N(Z)$ et $D(Z)$:

$$H_T(Z) = a_0 \cdot \frac{N(Z)}{D(Z)}$$

avec

$$N(Z) = 1 - 0{,}665\, Z^{-1} + Z^{-2}$$

$$D(Z) = 1 - 1{,}2146\, Z^{-1} + 0{,}6556\, Z^{-2}$$

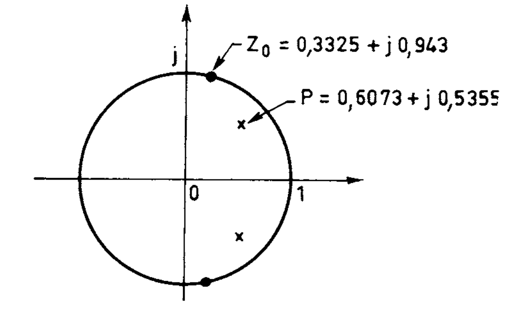


FIG. 6.8.a. *Pôles et zéros d'une cellule du second ordre générale*

La figure 6.8.b donne la réponse en fréquence du filtre. Les contributions du numérateur et du dénominateur de la fonction de transfert sont également représentées. Le facteur a_0 correspond à un facteur d'échelle qui est calculé pour que la réponse du filtre ait une valeur spécifiée à une fréquence donnée. Par exemple :

$$H_T(0) = 1 \quad \text{conduit à :} \quad a_0 = 0{,}33.$$

Les caractéristiques de temps de propagation de groupe et de phase sont données par les figures 6.5 et 6.6 respectivement.

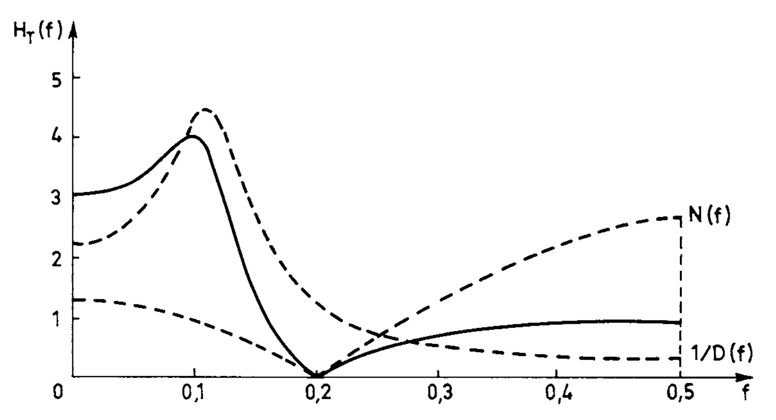


FIG. 6.8.b. *Réponse en fréquence de la cellule du second ordre générale*

La norme $\|H_T\|_2$ de la fonction $H_T(\omega)$ se calcule comme indiqué précédemment. Il vient :

$$\|H_T\|_2^2 = a_0^2 \frac{2 + a_1^2 + a_1^2 b_2 - 4a_1 b_1 + 2b_1^2 - 2b_2^2}{(1 - b_2)[(1 + b_2)^2 - b_1^2]} \tag{6.43}$$

Un cas particulier important du même type de filtre est le filtre dit « à encoche », utilisé pour retirer une fréquence pure d'un spectre sans perturber les autres composants. Sa fonction de transfert s'écrit :

$$H_E(Z) = \frac{1 + a_1 Z^{-1} + Z^{-2}}{1 + a_1(1 - \varepsilon)Z^{-1} + (1 - \varepsilon)^2 Z^{-2}} \tag{6.44}$$

où ε est un réel positif petit. Les pôles de cette cellule sont à la distance ε du cercle unité et des zéros comme le montre la figure 6.9.a. Pour ε très petit la bande d'affaiblissement à 3 dB, B_{3E} peut être approchée par :

$$B_{3E} \approx \frac{\varepsilon}{\pi}$$

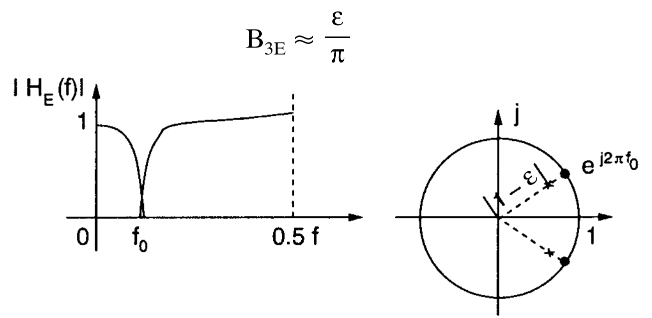


FIG. 6.9.a. *Filtre à encoche du second ordre*

En dehors de cette bande, les pôles et zéros se compensent et la réponse en fréquence est voisine de l'unité. De plus, un tel filtre apporte une très faible amplification sur un bruit blanc puisque, en appliquant (6.43), on obtient :

$$\|H_E\|_2^2 \approx \frac{2 - 3\varepsilon}{2 - 5\varepsilon} \approx 1 + \varepsilon$$

Si la fréquence du signal à éliminer n'est pas connue avec une grande précision ou si elle varie, alors il faut élargir la bande d'affaiblissement et écarter les zéros du cercle unité d'une quantité dont l'ordre de grandeur est donné par la relation (6.30).

La seconde catégorie de cellules du second ordre générales est celle des déphaseurs purs.

La cellule de circuit déphaseur est caractérisée par le fait que le numérateur et le dénominateur de la fonction de transfert ont les mêmes coefficients mais dans l'ordre inverse :

$$H_D(Z) = \frac{b_2 + b_1 Z^{-1} + Z^{-2}}{1 + b_1 Z^{-1} + b_2 Z^{-2}} = \frac{N(Z)}{D(Z)} \tag{6.45}$$

Les polynômes $N(Z)$ et $D(Z)$ sont des polynômes images. Il en résulte que $|H(e^{j\omega})| = 1$, c'est-à-dire que le circuit est un déphaseur pur.

En fonction des pôles et zéros la fonction de transfert $H_D(Z)$ s'écrit :

$$H_D(Z) = \frac{(P - Z^{-1})(\overline{P} - Z^{-1})}{(1 - PZ^{-1})(1 - \overline{P}Z^{-1})}$$

Il apparaît que les pôles et les zéros sont conjugués harmoniques. La figure 6.10 représente leur position dans le plan des Z.

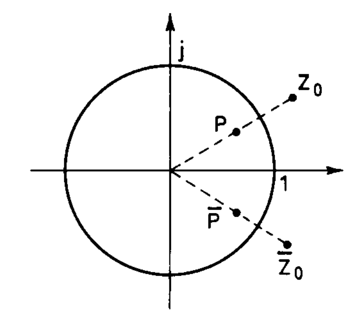


FIG. 6.10. *Pôles et zéros d'une cellule de déphaseur*

Le calcul de la phase et du temps de propagation de groupe de cette cellule se déduit très simplement des expressions (6.24) et (6.32) obtenues pour la cellule purement récursive. En effet on peut écrire :

$$H_D(Z) = \frac{N(Z)}{D(Z)} = \frac{Z^{-2} D(Z^{-1})}{D(Z)}$$

Comme d'autre part on a :

$$|D(\omega)| = |D(-\omega)|; \quad \varphi(\omega) = -\varphi(-\omega)$$

Il vient :

$$\varphi_D(\omega) = 2\varphi(\omega) + 2\omega$$

Le temps de propagation de groupe de la cellule de circuit déphaseur, $\tau_g(\omega)$, s'écrit tous calculs faits :

$$\tau_g(\omega) = \frac{1 - r^2}{1 - 2r\cos(\theta - \omega) + r^2} + \frac{1 - r^2}{1 - 2r\cos(\theta + \omega) + r^2} \qquad (6.46)$$

Il est facile de vérifier que, quand ω varie de 0 à π, la phase $\varphi_D(\omega)$ varie de 2π. En effet :

$$\varphi_D(\pi) = \int_0^\pi \tau_g(\omega)\, d\omega = 2\int_0^\pi \frac{1 - r^2}{1 + r^2 - 2r\cos\alpha}\, d\alpha = 2\pi$$

Une application intéressante de ce résultat est la possibilité de réaliser le filtre à encoche introduit précédemment à l'aide d'un circuit déphaseur, comme indiqué à la figure 6.9.b. Le zéro du filtre sur le cercle unité correspond à la fréquence où le déphasage est égal à π. En fait, c'est même un ensemble de deux filtres complémentaires qui sont obtenus avec un seul déphaseur du second ordre [1, 5].

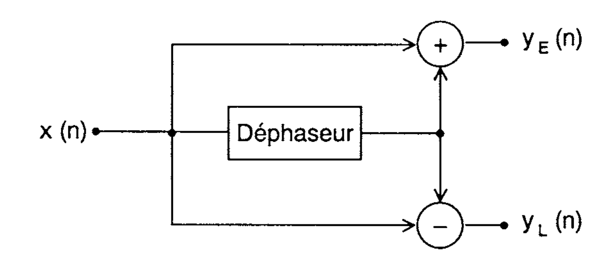


FIG. 6.9.b. *Réalisation d'un filtre à encoche et de son complément*

6.4 STRUCTURES POUR LA RÉALISATION

Les cellules sont réalisées par des circuits qui effectuent directement les opérations représentées dans l'expression des fonctions de transfert. Le terme Z^{-1} correspond à un retard d'une période élémentaire et est réalisé par une mise en mémoire ; les coefficients à utiliser dans les circuits sont ceux de la fonction de transfert avec le même signe pour le numérateur et le signe opposé pour le dénominateur.

Le circuit qui correspond directement à la relation de définition de la cellule du second ordre purement récursive est donné par la figure 6.11.

Les nombres de sortie $y(n)$ sont retardés deux fois, multipliés par les coefficients $-b_1$ et $-b_2$ avant d'être ajoutés aux nombres d'entrée $x(n)$. Le circuit comprend deux mémoires de données et deux mémoires de coefficients. Il faut effectuer, pour obtenir chaque nombre de sortie, deux multiplications et deux additions.

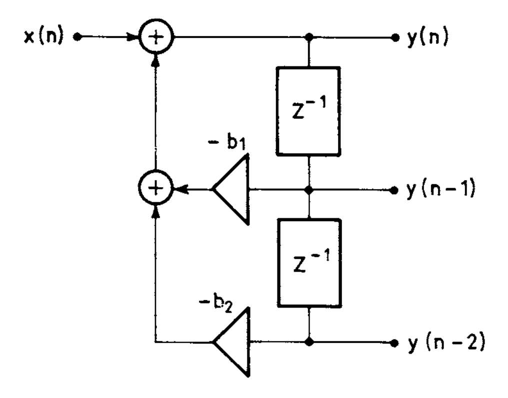


FIG. 6.11. *Circuits de la cellule purement récursive*

La cellule du second ordre générale peut être réalisée conformément à la relation de définition. Cependant, il faut alors deux mémoires de données pour les nombres d'entrée et deux mémoires pour les nombres de sortie. La structure obtenue n'est pas canonique, elle ne comporte pas le minimum d'éléments. En effet, il suffit de deux mémoires de données, si la fonction de transfert est factorisée comme suit :

$$H_T(Z) = \frac{N(Z)}{D(Z)} = \frac{1}{D(Z)} \cdot N(Z)$$

c'est-à-dire que les calculs correspondant au dénominateur sont effectués en premier et ceux qui correspondent au numérateur ensuite. La structure, dite D-N, est représentée sur la figure 6.12; elle correspond à l'introduction des deux variables internes $u_1(n)$ et $u_2(n)$ formant un vecteur d'état $U(n)$ à $N = 2$ dimensions. Le système est décrit par les équations suivantes :

$$\begin{aligned} u_1(n+1) &= x(n) - b_1 u_1(n) - b_2 u_2(n) \\ u_2(n+1) &= u_1(n) \\ y(n) &= a_0 x(n) - a_0 b_1 u_1(n) - a_0 b_2 u_2(n) + a_1 u_1(n) + a_2 u_2(n) \end{aligned}$$

Ou encore, sous forme matricielle, conformément à (4.34) :

$$U(n+1) = \begin{bmatrix} -b_1 & -b_2 \\ 1 & 0 \end{bmatrix} . U(n) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} x(n) \qquad (6.47)$$

$$y(n) = [-a_0 b_1 + a_1, -a_0 b_2 + a_2] . U(n) + a_0 x(n)$$

Cette représentation d'état conduit ainsi à une réalisation canonique ayant le nombre minimal de variables internes et par suite de mémoires.

D'après les résultats du paragraphe 4.6 il existe une structure duale correspondant aux variables internes $v_1(n)$ et $v_2(n)$ telles que :

$$\begin{bmatrix} v_1(n+1) \\ v_2(n+1) \end{bmatrix} = \begin{bmatrix} -b_1 & 1 \\ -b_2 & 0 \end{bmatrix} \begin{bmatrix} v_1(n) \\ v_2(n) \end{bmatrix} + \begin{bmatrix} -a_0 b_1 + a_1 \\ -a_0 b_2 + a_2 \end{bmatrix} x(n)$$

$$y(n) = v_1(n) + a_0 x(n)$$

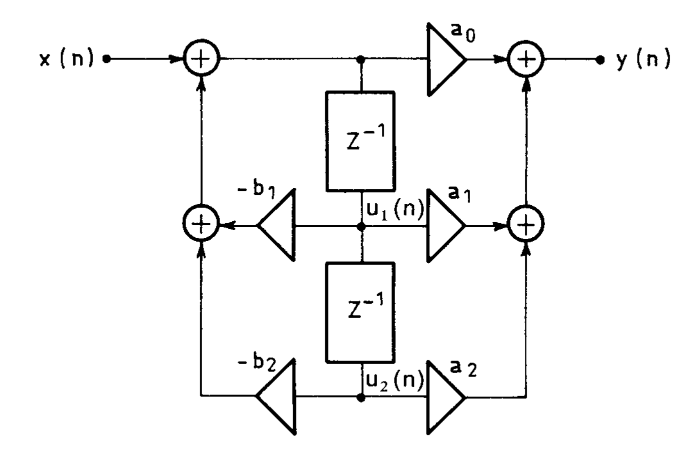


FIG. 6.12. *Cellule du second ordre en structure* D-N

Cette autre structure canonique est représentée sur la figure 6.13. Elle revient à effectuer d'abord les opérations du numérateur de la fonction de transfert en Z et est dite N-D.

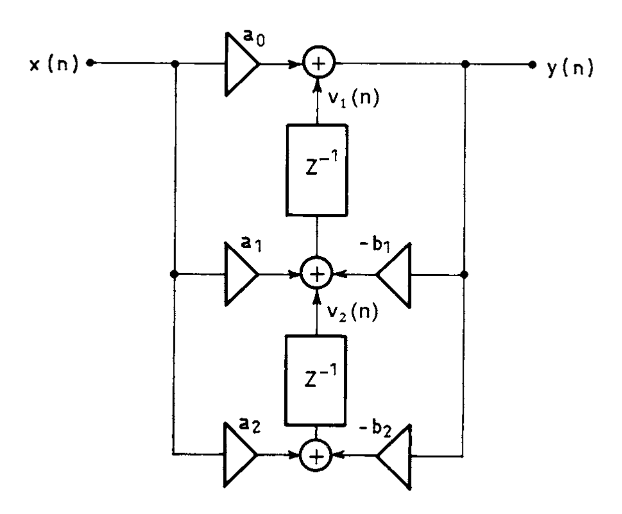


FIG. 6.13. *Cellule du second ordre en structure* N-D

La cellule du second ordre elliptique est généralement réalisée comme indiqué sur la figure 6.12-bis. Il faut effectuer quatre multiplications; celle qui porte sur le coefficient a_0, désigné par facteur d'échelle, est effectuée soit sur les nombres d'entrée $x(n)$, soit en sortie de la cellule comme sur la figure. Le gain en calcul par rapport aux schémas précédents apparaît clairement.

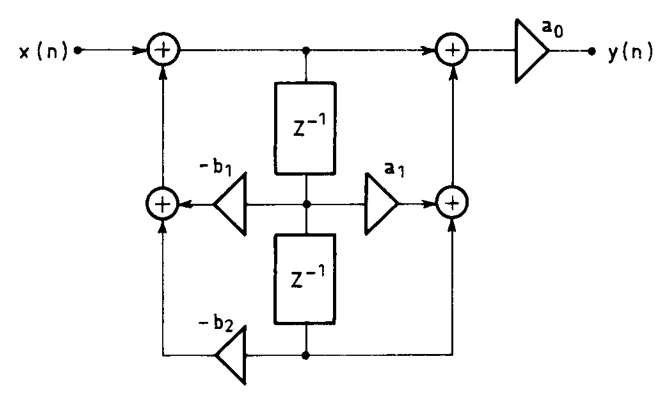


FIG. 6.12 bis. *Cellule du second ordre elliptique*

La cellule de déphaseur pur constitue un cas particulier pour la réalisation. En effet, la structure canonique ne permet pas d'exploiter les particularités de cette fonction : amplitude du signal constante et mêmes valeurs des coefficients au numérateur et au dénominateur. Une structure à 2 multiplications adaptée à cette fonction est donnée à la figure 6.14. La relation d'entrée-sortie correspondante s'écrit :

$$y(n) = x(n-2) + b_1[x(n-1) - y(n-1)] + b_2[x(n) - y(n-2)]$$

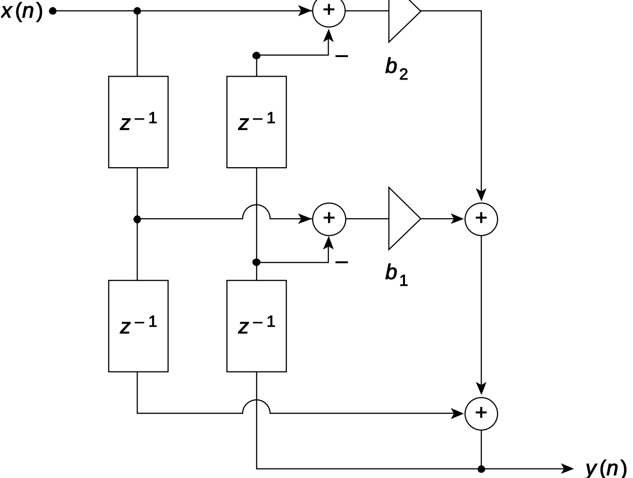


FIG. 6.14. *Cellule de déphaseur du second ordre à 2 multiplications*

Au total, il faut 2 multiplications, 4 additions et 4 mémoires. A noter que le déphaseur, par définition, n'a pas de facteur d'échelle, la figure 6.14 permet de ne mettre en mémoire que des signaux d'amplitude constante, ce qui minimise la longueur des mémoires et simplifie les estimations de bruit de calcul.

Pour optimiser la cellule et réduire le volume de circuits nécessaires par multiplication il est important de minimiser le nombre de bits de chacun des facteurs. Le cas des coefficients va être examiné en premier.

6.5 LIMITATIONS DU NOMBRE DE BITS DES COEFFICIENTS

La limitation du nombre de bits des coefficients se traduit par le fait qu'ils ne peuvent prendre qu'un nombre limité de valeurs; il s'en suit que les pôles ont un nombre limité de positions possibles à l'intérieur du cercle unité; il en est de même des zéros sur le cercle unité, dans le cas de la cellule de filtre elliptique. Ainsi la quantification à b_c bits de la valeur absolue des coefficients limite à 2^{2b_c} le nombre de positions que peuvent prendre les pôles dans un quart du cercle unité et 2^{b_c} le nombre de fréquences d'affaiblissement infini possible. La figure 6.15 représente ces positions pour $b_c = 3$.

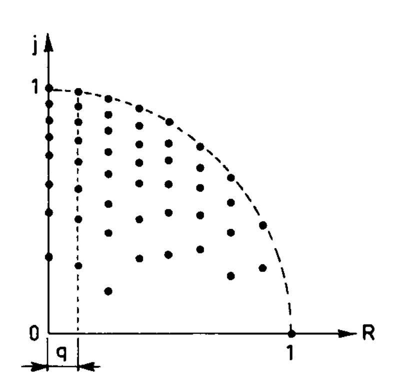


FIG. 6.15. *Positions des pôles et zéros avec une quantification à 3 bits des coefficients en valeur absolue*

Si la fonction de transfert de la cellule a été calculée d'abord et que la limitation du nombre de bits des coefficients intervient ensuite, la fonction de transfert H(Z) se trouve modifiée par l'introduction des polynômes parasites $e_N(Z)$ et $e_D(Z)$ au numérateur et au dénominateur [2]. On a la fonction $H_R(Z)$ telle que :

$$H_R(Z) = \frac{N(Z) + e_N(Z)}{D(Z) + e_D(Z)} \tag{6.49}$$

Si l'on désigne par δa_i et δb_i $(0 \leqslant i \leqslant 2)$ les erreurs d'arrondi faites sur les coefficients, les fonctions de transfert parasites s'écrivent :

$$e_N(f) = \sum_{i=0}^{2} \delta a_i e^{-j2\pi f i}; \quad e_D(f) = \sum_{i=1}^{2} \delta b_i e^{-j2\pi f i}$$

Considérons le cas de la cellule de filtre elliptique dont les coefficients sont quantifiés à b_c bits, signe compris, par arrondi. Compte tenu des inégalités :

$$|a_i| \leqslant 2; \quad |b_i| < 2$$

l'échelon de quantification q s'écrit :

$$q = 2 \times 2^{1-b_c} = 2^{2-b_c}$$

Il vient :

$$|e_D(f)| \leqslant 2 \cdot \frac{q}{2} = 2^{2-b_c} \quad (6.50)$$

Les modifications de la fonction de transfert dues à la quantification des coefficients du dénominateur prennent leur ampleur maximale pour les fréquences voisines des pôles, car alors la fonction $D(f)$ passe par un minimum.

En supposant qu'il n'y a pas d'arrondi des coefficients au numérateur, on peut écrire :

$$H_R(f) = \frac{N(f)}{D(f) + e_D(f)} \simeq \frac{N(f)}{D(f)} \left[1 - \frac{e_D(f)}{D(f)} \right]$$

L'erreur relative globale sur la réponse, $e(f) = \dfrac{H_R(f) - H(f)}{H(f)}$, se trouve alors bornée par :

$$|e_D(f)| \leqslant q \cdot \frac{1}{|D(f)|} \quad (6.51)$$

Cette expression permet de déterminer le nombre de bits nécessaire pour représenter les coefficients du dénominateur en fonction de la tolérance sur la réponse en fréquence et des valeurs des coefficients. Elle est utilisée au chapitre suivant.

Au numérateur la quantification du coefficient a_1 de la cellule de filtre elliptique amène une modification de l'abscisse des zéros qui se déplacent sur le cercle unité. Le déplacement de la pointe d'affaiblissement infini f_i qui en résulte prend la valeur df_i telle que :

$$|df_i| \leqslant \frac{1}{2\pi} \frac{2^{-b_c}}{|\sin 2\pi f_i|} \quad (6.52)$$

La quantification du coefficient a_0 de la cellule elliptique amène simplement une modification du gain du filtre.

6.6 LIMITATION DU NOMBRE DE BITS DES MÉMOIRES DE DONNÉES

Dans la structure D-N, qui est la plus couramment utilisée, le second facteur de la multiplication est le nombre contenu dans la mémoire de données. Cette mémoire a nécessairement une capacité limitée; la structure en boucle (fig. 6.12) fait que, même si les nombres d'entrée $x(n)$ ont un nombre de bits limité et que les

mémoires sont vides à la mise en fonctionnement, le nombre de bits des données à mettre en mémoire croît indéfiniment. Une limitation, en général par arrondi, est nécessaire. D'autre part les cellules peuvent présenter des gains importants et des instabilités, ce qui conduit à introduire une limitation de l'amplitude des données à mettre en mémoire avec un dispositif de saturation logique. Le schéma de la cellule elliptique du second ordre avec dispositif de saturation et quantification est donné par la figure 6.16 pour la structure D-N. Ce schéma suppose un seul dispositif de limitation situé juste avant la mise en mémoire, pour simplifier l'analyse du circuit.

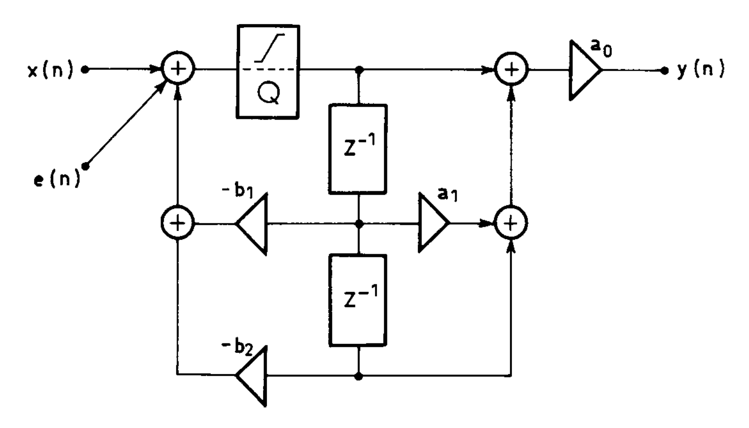


FIG. 6.16. *Cellule elliptique avec dispositif de limitation*

Le dispositif de quantification apporte une dégradation au signal qui traverse la cellule de filtre, c'est le bruit de calcul. En effet, suivant le schéma de la figure 6.16, la quantification revient à superposer au signal d'entrée $x(n)$ un signal d'erreur $e(n)$ qui traverse également le filtre. Si l'échelon de quantification a la valeur q, ce signal d'erreur est considéré comme ayant un spectre à répartition uniforme et une puissance $\frac{q^2}{12}$. Dans ces conditions le bruit de calcul B_c en sortie de la cellule a pour expression, si $fe = 1$ et d'après la relation (4.25) du paragraphe 4.4 :

$$B_c = \frac{q^2}{12} \int_0^1 \left| \frac{N(f)}{D(f)} \right|^2 df$$

ou encore en fonction de la suite $h(n)$, réponse impulsionnelle du filtre :

$$B_c = \frac{q^2}{12} \sum_{n=0}^{\infty} |h(n)|^2$$

En utilisant les résultats des paragraphes précédents, dans le cas d'une cellule du second ordre purement récursive à pôles complexes de coordonnées polaires (r, θ), c'est-à-dire la relation (6.36), il vient :

$$B_c = \frac{q^2}{12} \frac{1 + r^2}{1 - r^2} \frac{1}{1 + r^4 - 2r^2 \cos 2\theta} \tag{6.53}$$

et pour la cellule elliptique avec la relation (6.43) :

$$\mathrm{B}_c = \frac{q^2}{12} \frac{a_0^2(2 + a_1^2 + a_1^2 b_2 - 4a_1 b_1 + 2b_1^2 - 2b_2^2)}{(1 - b_2)\,[(1 + b_2)^2 - b_1^2]} \qquad (6.54)$$

L'échelon de quantification q est lié au nombre de bits des mémoires internes. La relation fait intervenir l'amplitude de la réponse en fréquence à la résonance de la partie purement récursive ; elle est étudiée en détail au chapitre suivant, dans la mise en cascade de cellules élémentaires.

Dans ce paragraphe seule la structure D-N a été considérée. Les calculs s'adaptent sans difficulté à la structure N-D [3].

L'introduction du dispositif de quantification a également des conséquences en l'absence de signal.

6.7 STABILITÉ ET AUTO-OSCILLATIONS

En l'absence de signal à l'entrée de la cellule de filtre RII, il se peut qu'un signal soit présent en sortie. C'est d'abord le cas si les coefficients sont tels que la cellule soit instable.

La condition de stabilité de la cellule est que les pôles soient à l'intérieur du cercle unité. Cette condition délimite un domaine de stabilité dans le plan (b_1, b_2). D'après les résultats du paragraphe 6.2 le domaine des pôles complexes est limité par la parabole :

$$b_2 = \frac{b_1^2}{4}$$

Alors la condition de stabilité impose :

$$0 \leqslant b_2 < 1$$

Si les pôles sont réels il faut de plus que :

$$-\frac{b_1}{2} + \frac{1}{2}\sqrt{b_1^2 - 4b_2} < 1\,; \quad -1 < -\frac{b_1}{2} - \frac{1}{2}\sqrt{b_1^2 - 4b_2}$$

ou encore :

$$b_2 > -b_1 - 1\,; \quad b_2 > -1 + b_1\,; \quad \text{soit} \quad |b_1| < 1 + b_2 \qquad (6.55)$$

Le domaine de stabilité correspondant est un triangle délimité par les trois droites :

$$b_2 = 1\,; \quad b_2 = -b_1 - 1\,; \quad b_2 = b_1 - 1$$

comme représenté sur la figure 6.17.

Cependant, même si la condition de stabilité est remplie, il se peut qu'en l'absence de signal à l'entrée, un signal soit présent en sortie du circuit ; c'est en général un signal constant ou périodique qui correspond à une auto-oscillation du filtre,

fréquemment appelé cycle limite. De telles auto-oscillations peuvent se produire aux grandes amplitudes par dépassement de la capacité des mémoires s'il n'y a pas de dispositif de saturation logique. L'équation du système en l'absence de signal d'entrée s'écrit :

$$y(n) + b_1 y(n-1) + b_2 y(n-2) = 0$$

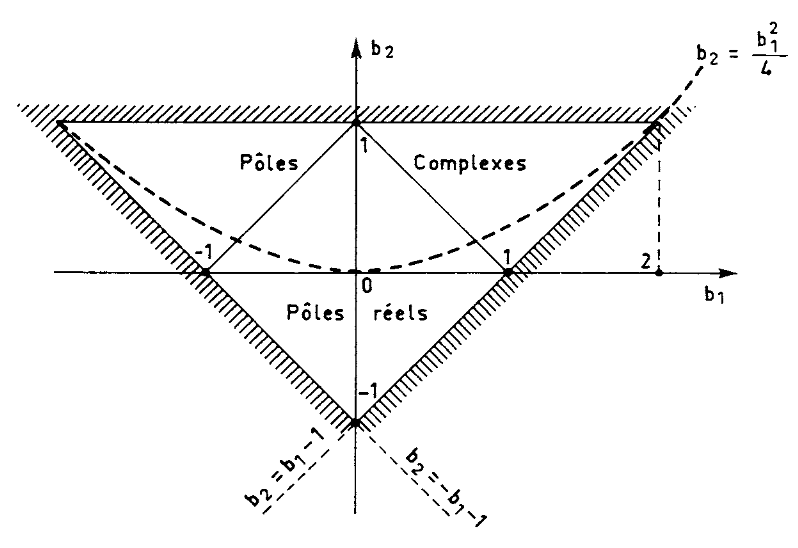


FIG. 6.17. *Domaine de stabilité de la cellule du second ordre*

La condition naturelle d'absence de telles oscillations s'exprime par l'inégalité :

$$|b_1 y(n-1) + b_2 y(n-2)| < 1$$

d'où la condition nécessaire et suffisante d'absence d'auto-oscillations aux grandes amplitudes :

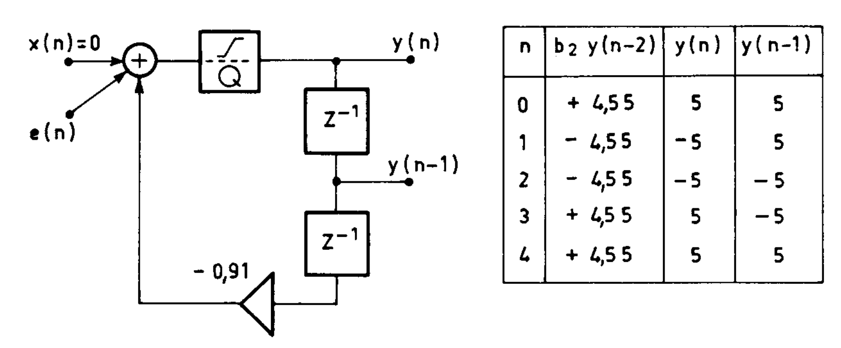
$$|b_1| + |b_2| < 1 \tag{6.56}$$

Cette inégalité délimite dans le plan (b_1, b_2) un carré à l'intérieur du triangle de stabilité de la cellule.

Pour éliminer toute possibilité d'oscillations de grande amplitude dues au dépassement de capacité des mémoires, on démontre qu'il suffit d'utiliser un dispositif de saturation logique comme indiqué au paragraphe 6.6 [4].

Des auto-oscillations se produisent également en raison de la quantification avant mise en mémoire. Elles sont de faibles amplitudes, dans les systèmes bien conçus. Elles tiennent au fait qu'en réalité le signal d'entrée n'est jamais nul, puisque même en l'absence de données $x(n)$, le signal d'erreur $e(n)$ dû à la quantification des nombres avant la mise en mémoire est appliqué au filtre.

Un exemple est donné par la figure 6.18.



n	$b_2\,y(n-2)$	y(n)	y(n−1)
0	+ 4,55	5	5
1	− 4,55	−5	5
2	− 4,55	−5	−5
3	+ 4,55	5	−5
4	+ 4,55	5	5

FIG. 6.18. *Auto-oscillations aux faibles amplitudes*

Une borne peut être obtenue simplement pour de tels signaux, compte tenu du fait que le signal d'erreur $e(n)$ est lui-même borné dans l'arrondi avec échelon q par :

$$|e(n)| \leqslant \frac{q}{2}$$

Si le filtre a comme réponse impulsionnelle la suite $h(i)$, les auto-oscillations sont limitées et il vient :

$$|y(n)| \leqslant \frac{q}{2} \sum_i |h(i)|$$

Cette borne est en fait très large; une estimation plus réaliste de l'amplitude A_a des auto-oscillations est donnée par l'expression :

$$A_a = \frac{q}{2} \operatorname{Max} |H(\omega)|$$

où $H(\omega)$ est la fonction de transfert de la cellule.

L'application à une cellule purement récursive du second ordre à pôles complexes conduit, d'après (6.37) et (6.29), à :

$$|y(n)| \leqslant \frac{q}{2} \frac{1}{(1-r)\sin\theta} \tag{6.57}$$

$$A_a = \frac{q}{2} \frac{1}{(1-r^2)\sin\theta} \tag{6.58}$$

Ces signaux ont souvent un spectre de raies correspondant à des fréquences voisines de celle où $H(\omega)$ est maximum et qui sont, soit diviseurs, soit dans un rapport simple avec la fréquence d'échantillonnage.

Dans la conception des filtres le nombre de bits des mémoires doit être suffisamment grand et l'échelon q suffisamment petit pour que ces auto-oscillations ne soient pas gênantes. Il faut noter également qu'elles peuvent être éliminées par l'utilisation d'un autre type de quantification que l'arrondi, par exemple la tronca-

tion de la valeur absolue [5]. C'est alors au prix d'une augmentation de la puissance de bruit en présence de signal.

Les résultats obtenus dans ce chapitre vont être utilisés dans le chapitre suivant qui traite de la mise en cascade de cellules élémentaires.

BIBLIOGRAPHIE

[1] P. A. Regalia, S. K. Mitra, P. P. Vaidyanathan – All-pass Filter : a Versatile Signal Processing Building Block. *Proceedings of the IEEE,* 76, N° 1, January 1988, pp. 19-37.

[2] J. B. Knowles and E. M. Olcayto – Coefficient Accuracy and Digital Filter Response. *IEEE Trans. on Circuit Theory,* March 1968.

[3] L. B. Jackson – On the Interaction of Round-off Noise and Dynamic Range in Digital Filters. *BSTJ*, Feb 1970.

[4] P. Ebert, J. Mazo and M. Taylor – Overflow Oscillations in Digital Filters. *BSTJ*, Nov. 1969.

[5] T. Claasen, W. Mecklenbrauker and J. Peek – Effects of Quantizations and Overflow in Recursive Digital Filters. *IEEE Trans.,* Vol. ASSP-24, N° 6, Dec 1976.

[6] S. K. Mitra and J. F. Kaiser – *Handbook for Digital Signal Processing,* John Wiley, New York, 1993.

EXERCICES

1 Soit à étudier la cellule du premier ordre :

$$y(n) = x(n) + by(n-1)$$

dans les conditions suivantes :

$$x(n) = 0; \qquad n < 0$$
$$x(n) = \cos n\omega; \quad n \geqslant 0$$
$$b = -0{,}8; \quad \omega = \frac{\pi}{2}; \quad y(-1) = 0$$

Donner l'expression de $y(n)$ en faisant apparaître les régimes transitoire et permanent.

Calculer l'amplitude et la phase de la réponse à partir de la valeur de $y(n)$; vérifier à l'aide des résultats du paragraphe 6.1. En considérant dans $y(n)$ le régime permanent, calculer le retard apporté par le filtre. Comparer la valeur obtenue au temps de propagation de groupe. Justifier la différence entre les deux valeurs.

2 Calculer la réponse du système défini par l'équation :

$$y(n) = x(n) + x(n-1) - 0{,}8y(n-1)$$

à la suite unitaire $u_0(n)$ et à la suite $x(n)$ telle que :

$$x(n) = 0 \quad \text{pour} \quad n < 0$$
$$x(n) = 1 \quad \text{pour} \quad n \geqslant 0$$

Donner la réponse en fréquence en régime permanent et transitoire.

3 Soit la cellule du second ordre purement récursive dont les coefficients sont :

$$b_1 = -1{,}56\,; \quad b_2 = 0{,}8$$

Donner la position des pôles. Calculer la réponse en fréquence, en phase et le temps de propagation de groupe. Comment sont modifiées ces fonctions si l'on ajoute deux zéros en j et $-j$. Dans ce cas donner le schéma de réalisation sous forme D-N et évaluer le volume de calculs à faire pour chaque nombre de sortie.

4 Donner l'expression de la réponse impulsionnelle de la cellule du second ordre purement récursive dont les coefficients sont :

$$b_1 = 1{,}60 \quad b_2 = 0{,}98$$

Calculer la fréquence de résonance et l'amplitude à la résonance. Exprimer la réponse $\mathrm{H}(\omega)$ et calculer la norme $\|\mathrm{H}\|_2$.

A cette cellule on adjoint deux zéros sur le cercle unité pour obtenir un affaiblissement infini à la fréquence $\dfrac{3fe}{2}$. Quels sont les coefficients de la cellule? Calculer la nouvelle expression de $\mathrm{H}(\omega)$ et la nouvelle valeur de $\|\mathrm{H}\|_2$?

Évaluer l'amplitude des auto-oscillations aux petits niveaux, de cette cellule, dans une réalisation sous forme D-N et sous forme N-D. Mettre en évidence un exemple d'auto-oscillation.

En l'absence de dispositif de saturation logique, cette cellule présente-t-elle des auto-oscillations de grande amplitude? Faire apparaître un exemple.

5 Avec combien de bits faut-il représenter les coefficients de la cellule dont la fonction de transfert en Z s'écrit :

$$\mathrm{H}(\mathrm{Z}) = \frac{1 - 0{,}952\,\mathrm{Z}^{-1} + \mathrm{Z}^{-2}}{1 - 1{,}406\,\mathrm{Z}^{-1} + 0{,}917\,z^{-2}}$$

pour que la réponse en fréquence ne soit pas modifiée de plus de 1 % au voisinage des pôles. Calculer le déplacement de la pointe d'affaiblissement infini.

6 Soit à réaliser un déphaseur pur du second ordre dont les pôles ont pour affixes $\mathrm{P}_{1,2}$ tels que :

$$\mathrm{P}_{1,2} = 0{,}71 \pm j\,0{,}54$$

Calculer les coefficients de la cellule et donner l'expression de la fonction $\tau_g(\omega)$. Montrer qu'il existe un schéma de réalisation qui conduit à un nombre de multiplications réduit. Cette cellule peut-elle présenter des auto-oscillations aux grandes amplitudes en l'absence de dispositif de saturation logique, et des auto-oscillations de faible amplitude?

Chapitre 7

Les filtres à réponse impulsionnelle infinie (RII)

Les filtres numériques à Réponse Impulsionnelle Infinie (RII), ou filtres récursifs ont des propriétés qui se rapprochent de celles des filtres analogiques et les techniques utilisées pour calculer leurs coefficients sont déduites de celles qui servent à déterminer les paramètres des filtres analogiques [1, 2, 3].

Avant d'aborder les méthodes de calcul des coefficients, il est utile de donner un certain nombre d'expressions générales pour les caractéristiques de ces filtres.

7.1 EXPRESSIONS GÉNÉRALES POUR LES CARACTÉRISTIQUES

Le filtre RII général est un système qui, à la suite de données $x(n)$ fait correspondre la suite $y(n)$ telle que :

$$y(n) = \sum_{l=0}^{L} a_l x(n-l) - \sum_{k=1}^{K} b_k y(n-k) \tag{7.1}$$

La fonction de transfert en Z de ce système s'écrit :

$$H(Z) = \frac{\sum_{l=0}^{L} a_l Z^{-l}}{1 + \sum_{k=1}^{K} b_k Z^{-k}} \tag{7.2}$$

C'est le quotient de deux polynômes en Z, qui sont souvent de même degré.

Les coefficients a_l et b_k étant des nombres réels, $H(Z)$ est un nombre complexe tel que :

$$\overline{H(Z)} = H(\overline{Z})$$

La réponse en fréquence du filtre s'écrit avec les mêmes conventions que dans les chapitres précédents :

$$H(\omega) = |H(\omega)| e^{-j\varphi(\omega)}$$

Le module et la phase s'expriment à partir de H(Z) par les expressions suivantes :

$$|H(\omega)|^2 = [H(Z)\,H(Z^{-1})]_{Z = e^{j\omega}} \tag{7.3}$$

Par élévation de H(ω) au carré et en utilisant (7.3), il vient :

$$\varphi(\omega) = -\frac{1}{2j}\ln\left[\frac{H(Z)}{H(Z^{-1})}\right]_{Z = e^{j\omega}} \tag{7.4}$$

En dérivant la fonction φ(Z) par rapport à la variable complexe Z, on obtient :

$$\frac{d\varphi}{dZ} = -\frac{1}{2j}\left[\frac{H'(Z)}{H(Z)} + \frac{1}{Z^2}\,\frac{H'(Z^{-1})}{H(Z^{-1})}\right]$$

Pour $Z = e^{j\omega}$ il vient :

$$\frac{d\varphi}{dZ} = -\frac{1}{jZ}\,Re\left[Z\,\frac{d}{dZ}\ln(H(Z))\right]$$

D'où l'expression du temps de propagation de groupe :

$$\tau(\omega) = \frac{d\varphi}{d\omega} = \frac{d\varphi}{dZ}\,jZ = -Re\left[Z\,\frac{d}{dZ}\ln(H(Z))\right]_{Z = e^{j\omega}} \tag{7.5}$$

Les équations (7.3), (7.4) et (7.5) constituent des expressions concises pour le calcul des principales caractéristiques des filtres RII d'ordre quelconque.

Exemple

Soit :

$$H(Z) = \frac{1}{D(Z)} = \frac{1}{1 + b_1 Z^{-1} + b_2 Z^{-2}}$$

On a :

$$\tau(\omega) = Re\left[Z.\frac{D'(Z)}{D(Z)}\right]_{Z = e^{j\omega}} = -Re\left[\frac{b_1 Z^{-1} + 2b_2 Z^{-2}}{1 + b_1 Z^{-1} + b_2 Z^{-2}}\right]_{Z = e^{j\omega}}$$

d'où l'expression :

$$\tau(\omega) = 1 - \frac{1 - b_2^2 + b_1(1 - b_2)\cos\omega}{1 + b_1^2 + b_2^2 + 2b_1(1 + b_2)\cos\omega + 2b_2\cos(2\omega)} \tag{7.6}$$

qui est équivalente à l'expression donnée au chapitre précédent quand les pôles sont complexes avec $b_1 = -2r\cos\theta$ et $b_2 = r^2$.

D'autres expressions des caractéristiques des filtres RII peuvent être obtenues à partir de l'expression de H(Z) en fonction de ses pôles et de ses zéros ; si le numérateur et le dénominateur de H(Z) ont le même degré N, et si N est pair, il vient :

$$H(Z) = a_0 \prod_{i=1}^{\frac{N}{2}} \frac{1 + a_1^i Z^{-1} + a_2^i Z^{-2}}{1 + b_1^i Z^{-1} + b_2^i Z^{-2}}$$

Le carré du module de la fonction de transfert est égal au produit des carrés du module des fonctions élémentaires ; la phase et le temps de propagation de groupe sont les sommes des contributions des cellules élémentaires.

$$|H(\omega)|^2 = a_0 \prod_{i=1}^{\frac{N}{2}} |H_i(\omega)|^2$$

$$\tau(\omega) = \sum_{i=1}^{\frac{N}{2}} \tau_i(\omega)$$

Les expressions générales pour les caractéristiques des filtres RII données ci-dessus sont utilisées dans le calcul des coefficients.

7.2 CALCUL DIRECT DES COEFFICIENTS PAR LES FONCTIONS MODÈLES

Une méthode directe pour calculer les coefficients d'un filtre RII consiste à faire appel à une fonction modèle, qui est une fonction réelle définie sur l'axe des fréquences.

Les fonctions modèles considérées sont des fonctions connues pour leurs propriétés de sélectivité, les fonctions de Butterworth, Bessel, Tchebycheff et les fonctions elliptiques. Elles sont également utilisées pour le calcul des filtres analogiques. Elles constituent un modèle pour le carré de la fonction de transfert à obtenir. Cependant un obstacle apparaît pour leur utilisation au calcul de filtres numériques car elles ne sont pas périodiques alors que la fonction à obtenir a la période f_e. Il faut donc établir une correspondance entre l'axe réel et l'intervalle $[0, f_e]$. Une telle correspondance est fournie par une transformation conforme dans le plan complexe qui doit posséder les propriétés suivantes :

– Transformer l'axe imaginaire en le cercle unité.

– Transformer une fraction rationnelle de la variable complexe s en une fraction rationnelle de la variable complexe Z.

– Conserver la stabilité.

Une première approche consiste à chercher à conserver pour le filtre numérique la réponse impulsionnelle du filtre analogique.

7.2.1 Invariance impulsionnelle

Soit le filtre analogique défini par l'équation :

$$y'_a(t) = by_a(t) + x(t) \tag{7.7}$$

Il a pour fonction de transfert :

$$\mathrm{H}(s) = \frac{1}{s-b} \tag{7.8}$$

et pour réponse impulsionnelle :

$$h(t) = e^{bt}; \quad t \geqslant 0 \tag{7.9}$$

Un échantillonnage de cette réponse avec la période T fournit la suite :

$$h(n\mathrm{T}) = e^{b\mathrm{T}\cdot n}; \quad n \geqslant 0 \tag{7.10}$$

qui a pour transformée en Z :

$$\mathrm{H}(z) = \frac{1}{1 - e^{b\mathrm{T}}z^{-1}} \tag{7.11}$$

Le pôle b du filtre analogique est devenu $e^{b\mathrm{T}}$ pour le filtre numérique. La méthode se généralise à un nombre quelconque de pôles.

Cette méthode très simple est utilisée, par exemple, dans la simulation des systèmes analogiques par des calculateurs numériques. Pour le calcul de filtres, elle présente un grave inconvénient, du fait du repliement de la réponse en fréquence. En effet, dans l'opération d'échantillonnage, la réponse en fréquence du filtre analogique se trouve repliée dans la bande utile du filtre numérique et les spécifications d'amplitude ne peuvent pas être conservées.

Une autre approche consiste à établir une correspondance directe entre le fonctionnement du filtre analogique et du filtre numérique.

En reprenant l'équation différentielle du filtre analogique, on peut écrire, à l'instant $n\mathrm{T}$:

$$y_a(n\mathrm{T}) = y_a(n\mathrm{T}-\mathrm{T}) + \int_0^{\mathrm{T}} y'_a(n\mathrm{T}-\mathrm{T}+\tau)\,d\tau \tag{7.12}$$

Le calcul de l'intégrale par la formule du trapèze conduit à :

$$y_a(n\mathrm{T}) - y_a(n\mathrm{T}-\mathrm{T}) = \frac{\mathrm{T}}{2}\,[y'_a(n\mathrm{T}) + y'_a(n\mathrm{T}-\mathrm{T})] \tag{7.13}$$

soit :

$$y_a(n\mathrm{T}) - y_a(n\mathrm{T}-\mathrm{T}) = \frac{\mathrm{T}}{2}\,[by_a(n\mathrm{T}) + x(n\mathrm{T}) + by_a(n\mathrm{T}-\mathrm{T}) + x(n\mathrm{T}-\mathrm{T})]$$

En prenant la transformée en Z des deux membres, on obtient la fonction de transfert en Z :

$$H(Z) = \frac{1}{\frac{2}{T}\frac{1-Z^{-1}}{1+Z^{-1}} - b} \tag{7.14}$$

qui fait apparaître une relation entre s et Z. La transformation homographique correspondante est couramment désignée par transformation bilinéaire.

7.2.2 La transformation bilinéaire

Soit la transformation qui au point du plan complexe d'affixe Z fait correspondre le point d'affixe s tel que :

$$s = \frac{2}{T}\frac{1-Z^{-1}}{1+Z^{-1}} \tag{7.15}$$

Cette transformation correspond à une approximation de l'exponentielle, puisque l'on a, pour les faibles valeurs de sT :

$$Z = \frac{1+\frac{T}{2}s}{1-\frac{T}{2}s} \simeq e^{sT} \tag{7.15bis}$$

A tout point du cercle unité, $Z = e^{j\omega T}$, correspond un point s tel que :

$$s = \frac{2}{T}\frac{1-e^{-j\omega T}}{1+e^{-j\omega T}} = j\frac{2}{T}\operatorname{tg}\frac{\omega T}{2} \tag{7.16}$$

Par suite au cercle unité correspond l'axe imaginaire.

La relation qui donne Z en fonction de s s'écrit :

$$Z = \frac{\frac{2}{T}+s}{\frac{2}{T}-s} \tag{7.17}$$

Une fraction rationnelle en s est transformée en une fraction rationnelle en Z avec la particularité que le numérateur et le dénominateur de la fonction de Z ont le même degré.

D'autre part si s a une partie réelle négative, Z est en module inférieur à 1, c'est-à-dire que la partie du plan complexe de la variable s qui est à gauche de l'axe imaginaire est transformée en l'intérieur du cercle unité ; cette propriété permet de conserver les caractéristiques de stabilité des systèmes.

Dans la définition de la transformation, le facteur $\frac{2}{T}$ est un facteur d'échelle, $T = \frac{1}{f_e}$ est la période d'échantillonnage du système numérique. Ce facteur

contrôle la déformation de l'axe des fréquences qui intervient quand la transformation bilinéaire est appliquée pour obtenir la fonction de transfert en Z d'un système numérique à partir d'une fonction complexe de s. En effet le système numérique obtenu, a une réponse en fréquence fonction de la variable f_N qui est liée aux valeurs de la fonction analogique initiale sur l'axe imaginaire $j.f_A$ par la relation (7.16) ci-dessus. Il vient :

$$\pi f_A = \frac{1}{T} \operatorname{tg}(\pi f_N T) \qquad (7.18)$$

La figure 7.1 illustre cette relation. On peut remarquer que pour les fréquences très faibles la déformation est négligeable, ce qui justifie le choix du facteur d'échelle $\frac{2}{T}$.

Fig. 7.1. *Déformation en fréquence apportée par la transformation bilinéaire*

Bien que d'autres transformations puissent aussi s'appliquer au calcul des coefficients des filtres numériques, la transformation bilinéaire est la plus utilisée. Il faut remarquer qu'elle permet de calculer les filtres numériques à partir de fonctions de transfert de filtres analogiques ou en utilisant les programmes de calcul mis au point pour les filtres analogiques.

Il faut prendre garde cependant que la réponse en fréquence se trouve déformée, comme indiqué précédemment, les pulsations analogique ω_A et numérique ω_N étant liées par la relation :

$$\omega_A = \frac{2}{T} \operatorname{tg}\left[\frac{\omega_N T}{2}\right] \qquad (7.19)$$

Le temps de propagation de groupe est également modifié :

$$\tau_N = \tau_A \left[1 + \left(\frac{\omega_A T}{2}\right)^2\right] \qquad (7.20)$$

c'est-à-dire qu'un filtre qui, en analogique, a un temps de groupe presque constant est transformé en un filtre numérique qui n'a pas cette propriété.

Pour calculer un filtre numérique à partir d'un gabarit selon cette méthode, il faut prédéformer le gabarit, calculer le filtre analogique satisfaisant au nouveau gabarit et ensuite appliquer la transformation bilinéaire.

7.2.3 Les filtres de Butterworth

Pour illustrer le calcul des coefficients par une fonction modèle, deux cas sont retenus, les fonctions de filtrage de Butterworth en raison de leur simplicité et les fonctions elliptiques qui sont les plus utilisées.

Une fonction de Butterworth d'ordre n est définie par l'expression :

$$|F(\omega)|^2 = \frac{1}{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}} \tag{7.21}$$

Le paramètre ω_c donne la valeur de la variable pour laquelle la fonction prend la valeur $\frac{1}{2}$. La figure 7.2 représente cette fonction pour diverses valeurs de n.

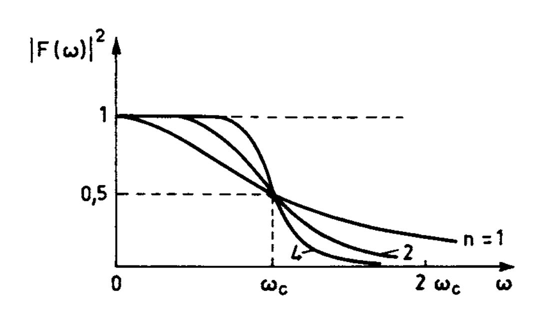


FIG. 7.2. *Fonctions de Butterworth*

Par extension analytique on peut écrire en prenant $\omega_c = 1$:

$$|F(\omega)|^2 = |H(j\omega)|^2 = |H(s)\,H(-s)|_{s=j\omega} = \frac{1}{1+\omega^{2n}}$$

$$H(s)\,H(-s) = \frac{1}{1 + \left(\frac{s}{j}\right)^{2n}} = \frac{1}{1 + (-s^2)^n}$$

Les pôles de cette fonction sont sur le cercle unité ; par exemple pour n impair on peut écrire :

$$H(s)\,H(-s) = \frac{1}{\prod_{k=1}^{2n} \left(s - e^{j\pi\frac{k}{n}}\right)}$$

En affectant à H(s) les pôles qui sont à gauche de l'axe imaginaire afin d'aboutir à un filtre stable, et après regroupement pour obtenir des cellules du premier ou de second ordre à coefficients réels, il vient :

$$H(s) = \frac{1}{1+s} \prod_{k=1}^{\frac{n-1}{2}} \frac{1}{s^2 + 2\cos\left(\pi \frac{k}{n}\right).s + 1}$$

De la même manière on obtient pour n pair :

$$H(s) = \prod_{k=1}^{\frac{n}{2}} \frac{1}{s^2 + 2\cos\left[\frac{\pi(2k-1)}{2n}\right].s + 1}$$

Le filtre numérique correspondant est fourni par le changement de variable (7.15), avec un facteur d'échelle convenable, c'est-à-dire :

$$s = \frac{1}{\operatorname{tg}\frac{\omega_c T}{2}} \cdot \frac{1 - Z^{-1}}{1 + Z^{-1}}$$

Le point de l'axe des fréquences où la réponse du filtre numérique prend la valeur $\frac{1}{\sqrt{2}}$ est bien f_c tel que $\omega_c = 2\pi f_c$.

En posant : $u = \frac{1}{\operatorname{tg}(\pi f_c T)}$ et $\alpha_k = 2\cos\left(\frac{\pi k}{n}\right)$ on obtient la fonction de transfert en Z suivante pour le filtre numérique d'ordre n impair :

$$H(Z) = \frac{1 + Z^{-1}}{(1+u) + (1-u)Z^{-1}} \prod_{k=1}^{\frac{n-1}{2}} a_0^k \frac{(1 + Z^{-1})^2}{1 + b_1^k Z^{-1} + b_2^k Z^{-2}}$$

avec :

$$a_0^k = \frac{1}{1 + u\alpha_k + u^2}\ ; \quad b_1^k = 2a_0^k(1 - u^2)\ ; \quad b_2^k = a_0^k(1 - u\alpha_k + u^2)$$

Pour n pair, avec $\alpha_k = 2\cos\left(\pi \frac{2k-1}{2n}\right)$, il vient :

$$H(Z) = \prod_{k=1}^{\frac{n}{2}} a_0^k \frac{(1 + Z^{-1})^2}{1 + b_1^k Z^{-1} + b_2^k Z^{-2}} \tag{7.22}$$

Il apparaît ainsi que les zéros de la fonction de transfert en Z se trouvent tous au point Z = – 1, ce qui peut simplifier la réalisation du filtre. D'autre part la fonction est complètement déterminée par la donnée des deux paramètres n et u.

L'ordre n se calcule à partir du gabarit du filtre. Soit à réaliser un filtre dont la réponse en fréquence soit supérieure ou égale à $1 - \delta_1$ dans la bande $[0, f_1]$ et inférieure ou égale à δ_2 dans la bande $\left[f_2, \frac{f_e}{2}\right]$. En revenant à la fonction modèle $F(\omega)$, ces contraintes impliquent les inégalités suivantes :

$$\frac{1}{1+\left(\frac{\omega_1}{\omega_c}\right)^{2n}} \geqslant (1-\delta_1)^2 \quad \text{et} \quad \frac{1}{1+\left(\frac{\omega_2}{\omega_c}\right)^{2n}} \leqslant \delta_2^2$$

Pour δ_1 et δ_2 petits, on obtient l'expression de n suivante :

$$n \geqslant \frac{\frac{1}{2}\log(2\delta_1) + \log(\delta_2)}{\log(\omega_1) - \log(\omega_2)}$$

ce qui donne pour l'ordre N du filtre numérique correspondant :

$$N \geqslant \frac{\log\left[\frac{1}{\delta_2\sqrt{2\delta_1}}\right]}{\log[\mathrm{tg}(\pi f_2 T)] - \log[\mathrm{tg}(\pi f_1 T)]} \tag{7.23}$$

Une fois n choisi le paramètre u doit être pris dans l'intervalle :

$$\frac{1}{\mathrm{tg}(\pi f_2 T)}\left(\frac{1}{\delta_2}\right)^{\frac{1}{n}} \leqslant u \leqslant \frac{(2\delta_1)^{\frac{1}{2n}}}{\mathrm{tg}(\pi f_1 T)}$$

Ces paramètres permettent de calculer $H(Z)$.

Exemple

Soit le gabarit suivant, déjà considéré pour les filtres RIF ;

$$\delta_1 = 0{,}045; \quad \delta_2 = 0{,}015; \quad f_e = 1; \quad f_1 = 0{,}1725; \quad f_2 = 0{,}2875$$

on trouve $N \simeq 7{,}3$; la valeur retenue est $N = 8$ et pour le choix minimal de u il vient :

$$H(Z) = 0{,}00185\ \frac{(1+Z^{-1})^2}{1-0{,}36\,Z^{-1}+0{,}04\,Z^{-2}} \cdot \frac{(1+Z^{-1})^2}{1-0{,}39\,Z^{-1}+0{,}12\,Z^{-2}} \cdot \frac{(1+Z^{-1})^2}{1-0{,}45\,Z^{-1}+0{,}31\,Z^{-2}} \cdot \frac{(1+Z^{-1})^2}{1-0{,}58\,Z^{-1}+0{,}69\,Z^{-2}}$$

La réponse en fréquence obtenue est donnée par la figure 7.3 et le temps de propagation de groupe par la figure 7.4.

Quand la bande de transition du filtre, $\Delta f = f_2 - f_1$ est suffisamment faible, l'expression (7.23) se simplifie comme suit :

$$N \approx \log\left[\frac{1}{\delta_2\sqrt{2\delta_1}}\right] \cdot \frac{2{,}3}{2\pi} \cdot \frac{f_e}{\Delta f} \cdot \sin(2\pi f_1/f_e) \tag{7.24}$$

Ainsi l'ordre du filtre est alors inversement proportionnel à la largeur de la bande de transition, comme pour les filtres RIF. Il s'en suit que la sélectivité reste assez limitée en pratique.

Finalement les filtres de Butterworth sont faciles à calculer, des simplifications importantes peuvent apparaître dans leur réalisation en raison de la configuration des zéros et aussi dans certains cas des pôles, mais leur sélectivité est beaucoup plus faible que celle du filtre optimal que constitue le filtre elliptique.

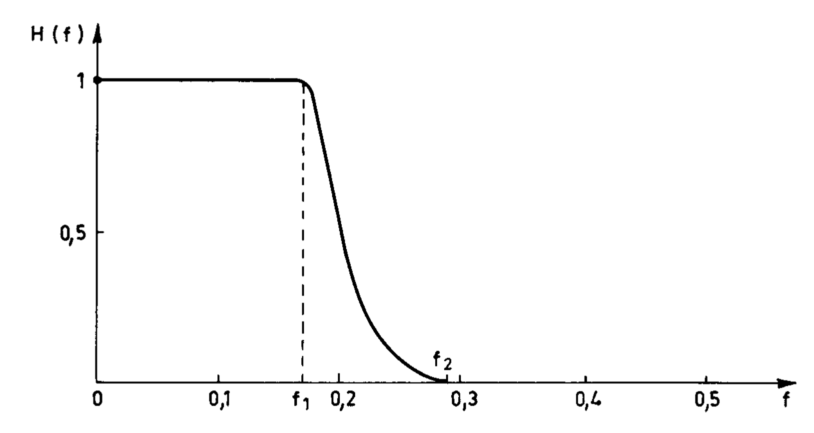


FIG. 7.3. *Réponse en fréquence d'un filtre de Butterworth d'ordre 8*

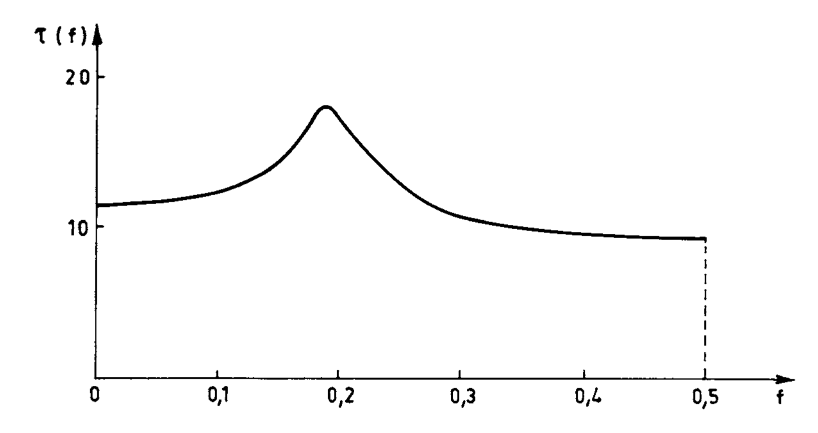


FIG. 7.4. *Temps de propagation de groupe du filtre de Butterworth*

7.2.4 Les filtres elliptiques

Le filtre elliptique présente des ondulations en bande passante et en bande affaiblie. Il est optimal en ce sens que pour un ordre n donné et des amplitudes d'ondu-

lations fixées il présente la bande de transition la plus faible. La fonction modèle fait appel aux fonctions elliptiques ; elle s'écrit :

$$T^2(u) = \frac{1}{1 + \varepsilon^2 sn^2(u, k_1)} \tag{7.25}$$

où $y = sn(u, k)$ est défini de façon implicite par la fonction elliptique incomplète de première espèce :

$$u = \int_0^{\text{Arc}\sin y} \frac{d\theta}{(1 - k^2 \sin^2 \theta)^{\frac{1}{2}}} \tag{7.26}$$

Une représentation de la fonction $T^2(u)$ pour $u = j\omega$ est donnée par la figure 7.5 où sont indiqués les paramètres correspondant à k_1 tel que :

$$k_1 = \frac{\varepsilon}{\sqrt{A^2 - 1}}$$

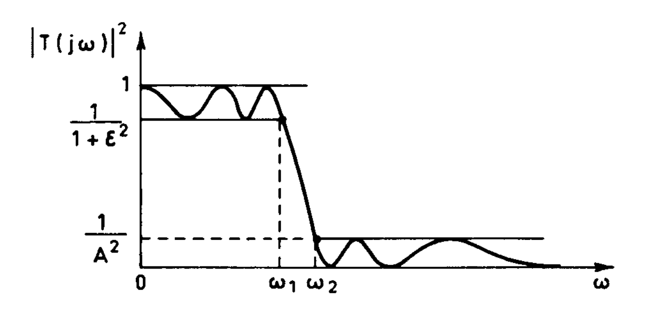


FIG. 7.5. *Fonction de filtrage elliptique*

La fonction $sn^2(\omega, k)$ oscille entre 0 et 1 pour $\omega < \omega_1$ et entre $\frac{\sqrt{A^2-1}}{\varepsilon}$ et l'infini pour $\omega \geqslant \omega_2$.

On montre que l'ordre n du filtre est déterminé à partir des paramètres k_1 et k, facteur de sélectivité, tel que :

$$k = \frac{\omega_1}{\omega_2}$$

par l'expression :

$$n = \frac{K(k)K(\sqrt{1 - k_1^2})}{K(k_1)K(\sqrt{1 - k^2})} \tag{7.27}$$

où K(k) est l'intégrale elliptique complète de première espèce :

$$K(k) = \int_0^{\frac{\pi}{2}} \frac{d\theta}{(1 - k^2 \sin^2 \theta)^{\frac{1}{2}}} \qquad (7.27\text{-}bis)$$

Cette intégrale se calcule par la méthode d'approximation du polynôme de Tchebycheff qui conduit à une erreur de l'ordre de 10^{-8} avec un polynôme de degré 4. La fonction inverse de l'intégrale incomplète de première espèce est calculée par le quotient de deux séries rapidement convergentes.

Une expression simplifiée peut être obtenue pour l'ordre n du filtre à partir du gabarit général donné par la figure 5.7. Avec l'hypothèse d'une ondulation en bande passante comprise entre 1 et $1 - 2\delta_1$, et les paramètres suivants (fig. 7.5) :

$$\delta_2 = \frac{1}{A}; \quad 2\delta_1 = 1 - \frac{1}{\sqrt{1+\varepsilon^2}}; \quad f_e = 1$$

il vient :

$$n \simeq \frac{2}{\pi^2} \cdot \ln\left(\frac{2}{\delta_2\sqrt{\delta_1}}\right) \cdot \ln\left[\frac{8\,\omega_1}{\omega_2 - \omega_1}\right] \qquad (7.28)$$

L'ordre N du filtre numérique satisfaisant le gabarit de la figure 5.7 est alors donné par :

$$N \simeq \frac{2}{\pi^2} \cdot \ln\left(\frac{2}{\delta_2\sqrt{\delta_1}}\right) \cdot \ln\left[\frac{8\,\mathrm{tg}\,(\pi f_1/f_e)}{\mathrm{tg}\left(\pi\frac{f_2}{f_e}\right) - \mathrm{tg}\left(\pi\frac{f_1}{f_e}\right)}\right]$$

Généralement la bande de transition $\Delta f = f_2 - f_1$ est petite, et il vient, en passant aux logarithmes décimaux :

$$N \simeq 1{,}076 \cdot \log\left[\frac{2}{\delta_2\sqrt{\delta_1}}\right] \cdot \log\left[\frac{f_e}{\Delta f} \cdot \frac{4}{\pi} \cdot \sin\left(2\pi\frac{f_1}{f_e}\right)\right] \qquad (7.29)$$

Cette expression est à rapprocher de la relation (V.32) pour les filtres à réponse impulsionnelle finie. Ainsi pour les filtres RII de type elliptique, l'ordre est proportionnel au logarithme de l'inverse de la bande de transition normalisée, ce qui conduit à des valeurs beaucoup plus faibles que dans le cas des filtres RIF. De plus, la relation (7.29) montre que la largeur de bande intervient : la valeur maximale de N est atteinte pour f_1 voisin de $f_e/4$, c'est-à-dire une bande passante représentant la moitié de la bande utile. Une simplification supplémentaire peut être obtenue pour les filtres à bande passante étroite ; dans ce cas l'ordre N′ du filtre est donné par :

$$N' \simeq 1{,}076 \cdot \log\left[\frac{2}{\delta_2\sqrt{\delta_1}}\right] \cdot \log\left(8 \cdot \frac{f_1}{\Delta f}\right) \qquad (7.30)$$

et c'est la raideur de coupure qui intervient, comme pour les filtres analogiques.

Une fois l'ordre du filtre déterminé, la procédure de calcul comporte la détermination des pôles et des zéros de $T^2(u)$ qui présentent une double périodicité dans le plan complexe. Par changement de variable, puis application de la transformée bilinéaire on aboutit à la configuration des pôles et zéros du filtre numérique dans le plan des Z [4].

Pour le calcul d'un filtre selon cette technique le gabarit est spécifié par :

– L'amplitude crête à crête des ondulations en bande passante exprimée en dB :

$$BP = -20 \log (1 - 2\delta_1)$$

– L'amplitude des ondulations en bande affaiblie exprimée en dB par :

$$AT = 20 \log \left(\frac{1}{\delta_2} \right)$$

– La fréquence de fin de bande passante : FB.
– La fréquence de début de bande affaiblie : FA.
– La fréquence d'échantillonnage : FE.

Exemple

Soit le gabarit donné au paragraphe précédent; BP = 0,4; AT = 36,5; FE = 1; FB = 0,1725; FA = 0,2875. On trouve N = 3.3, la valeur retenue est N = 4. Les zéros et les pôles ont les affixes suivants (fig. 7.6) :

$$E_1 = -0{,}816 + j\,0{,}578 \quad Z_2 = -0{,}2987 + j\,0{,}954$$
$$P_1 = -0{,}407 + j\,0{,}313 \quad P_2 = 0{,}335 + j\,0{,}776$$

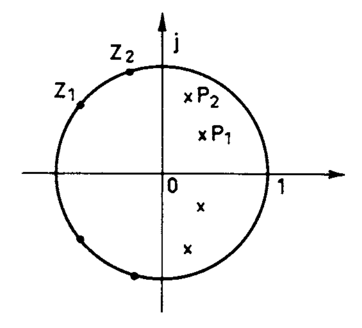


FIG. 7.6. *Pôles et zéros d'un filtre elliptique*

Pour faire apparaître les pointes d'affaiblissement infini, la courbe $\frac{1}{|H(f)|}$ donnant l'affaiblissement du filtre en fonction de la fréquence est représentée sur la figure 7.7.

La figure 7.8 donne le temps de groupe du filtre obtenu. Les courbes sont à comparer aux résultats obtenus avec le filtre de Butterworth calculé sur le même gabarit; elles font apparaître un avantage important pour le filtre elliptique qui demande un ordre deux fois plus faible et apporte une réduction de complexité des circuits dans la même proportion.

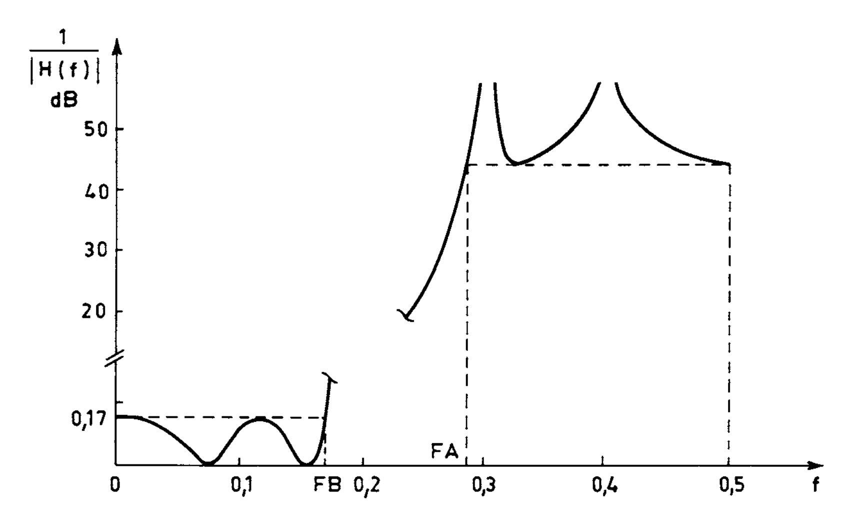


Fig. 7.7. *Réponse en fréquence d'un filtre elliptique d'ordre 4*

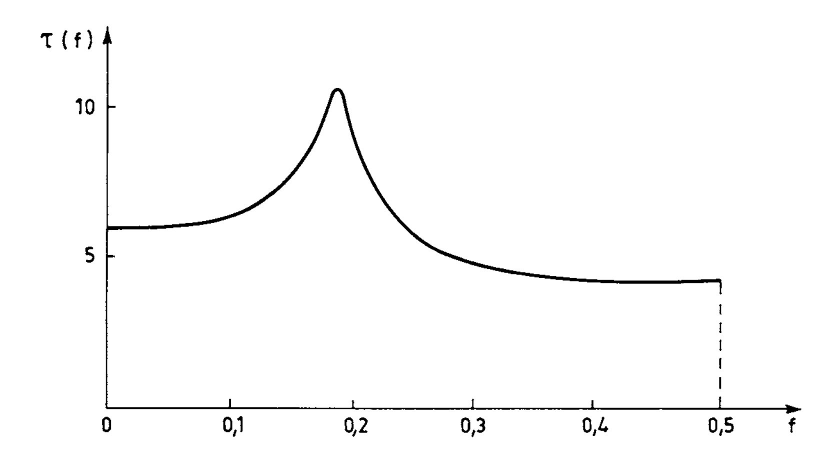


Fig. 7.8. *Temps de groupe d'un filtre elliptique d'ordre 4*

Les méthodes qui ont été exposées permettent de calculer des filtres passe-bas, à partir desquels il est possible d'obtenir des filtres passe-haut et passe-bande par des transformations en fréquence.

7.2.5 Calcul d'un filtre quelconque par transformation d'un passe-bas

Les procédures de calcul présentées dans les paragraphes précédents conduisent à une fonction $\mathrm{H}(s)$ qui par transformation bilinéaire fournit la fonction de transfert en Z du filtre numérique. Pour $s = j\omega$ la fonction $\mathrm{H}(\omega)$ est une fonction de filtrage

passe-bas sur le domaine des fréquences qui s'étend de zéro à l'infini. Il est possible de lui appliquer des transformations qui conduisent à d'autres types de filtres [5]. Par exemple pour obtenir une fonction passe-bas dont la bande passante s'étend de 0 à ω_1' à partir d'une fonction dont la bande passante couvre le domaine $[0, \omega_1]$ il faut faire la transformation :

$$s \to s\,\frac{\omega_1}{\omega_1'}$$

En partant d'un filtre passe-bas dont la bande passante est limitée à 1 on peut obtenir les filtres suivants, en désignant par ω_B et ω_H les limites inférieures et supérieures de la bande passante :

– Autre passe-bas : $s \to \dfrac{s}{\omega_H}$

– Passe-haut : $s \to \dfrac{\omega_B}{s}$

– Passe-bande : $s \to \dfrac{s^2 + \omega_H\omega_B}{s(\omega_H - \omega_B)}$

– Coupe bande : $s \to \dfrac{s(\omega_H - \omega_B)}{s^2 + \omega_H\omega_B}$

Ces transformations conservent les amplitudes des ondulations de la réponse du filtre mais amènent une déformation en fréquence.

Une méthode plus directe consiste à utiliser d'autres transformations que la transformation bilinéaire pour aboutir à l'expression H(Z) ; par exemple la transformation suivante permet d'obtenir directement un filtre numérique passe-bande à partir d'une fonction de filtrage en s passe-bas :

$$s = \frac{1}{T}\,\frac{1 - 2\cos(\omega_0 T)\,Z^{-1} + Z^{-2}}{1 - Z^{-2}} \tag{7.31}$$

Pour $Z = e^{j\omega T}$, il vient :

$$s = j\,\frac{1}{T}\,\frac{\cos(\omega_0 T) - \cos(\omega T)}{\sin(\omega T)} \tag{7.32}$$

Si la bande passante du filtre numérique s'étend de ω_B à ω_H, il faut choisir ω_0, tel que les abscisses des points transformés soient égales en valeur absolue et de signe opposé :

$$\frac{\cos(\omega_0 T) - \cos(\omega_B T)}{\sin(\omega_B T)} = -\,\frac{\cos(\omega_0 T) - \cos(\omega_H T)}{\sin(\omega_H T)}$$

d'où :

$$\cos(\omega_0 T) = \frac{\cos\left[(\omega_B + \omega_H)\dfrac{T}{2}\right]}{\cos\left[(\omega_B - \omega_H)\dfrac{T}{2}\right]}$$

Cette approche évite d'ajouter une étape dans la procédure de calcul d'un filtre passe-bande.

Il est également possible de faire appel à des transformations dans le plan des Z qui conservent le cercle unité ; la plus simple est la transformation de Z en – Z qui change un filtre passe-bas en passe-haut.

La transformation suivante où α est un nombre réel :

$$Z^{-1} \rightarrow \frac{Z^{-1} - \alpha}{1 - \alpha Z^{-1}} \tag{7.33}$$

change un passe-bas en un autre passe-bas. En fait on montre que la transformation la plus générale a pour expression [5] :

$$Z^{-1} \rightarrow \pm \prod_{k=1}^{K} \frac{Z^{-1} - \alpha_k}{1 - \alpha_k Z^{-1}} \tag{7.34}$$

avec $|\alpha_k| < 1$ pour assurer la stabilité.

Par exemple un passe-bas est transformé en passe-bande par :

$$Z^{-1} \rightarrow - \frac{Z^{-2} - \alpha_1 Z^{-1} + \alpha_2}{\alpha_2 Z^{-2} - \alpha_1 Z^{-1} + 1} \tag{7.35}$$

Il apparaît ainsi que l'on peut calculer d'une manière directe à partir de fonctions modèles tous les types de filtres. Cependant des limitations importantes interviennent d'abord parce que les ondulations du filtre doivent être constantes, par exemple pour les filtres elliptiques, dans les bandes passantes et affaiblies ; ensuite les méthodes exposées ne permettent pas de tenir compte d'éventuelles contraintes sur la réponse impulsionnelle. Pour lever ces limitations il faut faire appel aux techniques d'optimisation.

7.3 TECHNIQUES ITÉRATIVES POUR LE CALCUL DES FILTRES RII AVEC DES SPÉCIFICATIONS EN FRÉQUENCE

Les méthodes d'optimisation permettent, comme pour les filtres RIF, le calcul d'un filtre RII sur des spécifications quelconques. Cependant le calcul est un peu plus délicat dans ce cas car des précautions doivent être prises pour éviter d'aboutir à un système instable.

Deux méthodes vont être présentées qui correspondent à deux critères d'optimisation différents. Le premier critère est la minimisation de l'erreur quadratique [6].

7.3.1 Minimisation de l'erreur quadratique

La fonction de transfert d'un filtre est donnée sous forme factorisée par l'expression introduite précédemment :

$$H(Z) = a_0 \prod_{i=1}^{\frac{N}{2}} \frac{1 + a_1^i Z^{-1} + a_2^i Z^{-2}}{1 + b_1^i Z^{-1} + b_2^i Z^{-2}}\,; \quad a_0 > 0 \tag{7.36}$$

en considérant que le numérateur et le dénominateur sont de même degré N pair.

Soit $D(f)$ la fonction à approcher par la réponse en fréquence du filtre $H(f)$; l'écart entre ces fonctions représente une erreur qu'il est possible de minimiser au sens des moindres carrés, en un nombre de points, égal à N_0, de l'axe des fréquences; il vient alors :

$$E = \sum_{n=0}^{N_0-1} (|H(f_n)| - |D(f_n)|)^2$$

La valeur E est fonction d'un ensemble de 2N + 1 paramètres, qui sont les coefficients du filtre :

$$E = E(a_0, a_1^i, a_2^i, b_1^i, b_2^i) \quad \text{avec} \quad 1 \leqslant i \leqslant \frac{N}{2}$$

Le minimum correspond à l'ensemble des valeurs des 2N + 1 paramètres x_k tel que :

$$\frac{\partial E}{\partial a_k} = 0\,; \quad 1 \leqslant k \leqslant 2N+1$$

Pour le paramètre a_0 il vient en posant $H(Z) = a_0 H_1(Z)$

$$\frac{\partial E}{\partial x_k} = 0 = 2 \sum_{n=0}^{N_0-1} (a_0|H_1(f_n)| - |D(f_n)|)H_1(f_n)|$$

D'où la valeur de a_0 :

$$a_0 = \frac{\sum_{n=0}^{N_0-1} |D(f_n)|\,|H_1(f_n)|}{\sum_{n=0}^{N_0-1} |H_1(f_n)|^2} \tag{7.37}$$

Le problème d'optimisation se trouve ramené à 2N variables.

La procédure consiste à partir d'une fonction $H_1^0(Z)$ initiale, fournie par exemple par la méthode de calcul direct des filtres elliptiques donnée au paragraphe précédent, et à supposer que l'on se trouve suffisamment près de l'optimum pour que la fonction E puisse être assimilée à une fonction quadratique des 2N

paramètres x_k. Alors l'optimum cherché est fourni par un accroissement des paramètres donné par le vecteur à 2N éléments ΔX tel que :

$$E(X + \Delta X) \simeq E(X) + \sum_{k=1}^{2N} \frac{\partial E}{\partial x_k} \Delta x_k + \frac{1}{2} \sum_{k=1}^{2N} \sum_{l=1}^{2N} \frac{\partial^2 E}{\partial x_k \, \partial x_l} \Delta x_k \, \Delta x_l$$

En désignant par A la matrice à 2N lignes et à N_0 colonnes qui a pour éléments :

$$a_{ij} = 2 \frac{\partial}{\partial x_i} [a_0 . |H_1(f_j)|]$$

et par Δ le vecteur colonne à N_0 termes e_n, tels que :

$$e_n = a_0 . |H_1(f_n)| - |D(f_n)|$$

La condition des moindres carrés est obtenue en écrivant que E(X + ΔX) est extrémum. Comme au paragraphe 5.4 pour le calcul des coefficients des filtres RIF, il vient :

$$\Delta X = -[AA^t]^{-1} . A . \Delta$$

La méthode consiste ensuite à réitérer le calcul avec les nouvelles valeurs des paramètres, ce qui doit conduire à l'optimum cherché. Les chances d'atteindre ce but et la rapidité de la méthode dépendent des accroissements donnés aux paramètres; la meilleure stratégie est sans doute celle qui est fournie par l'algorithme de Fletcher et Powell [7].

Pour s'assurer de la stabilité du système obtenu on peut contrôler la stabilité à chaque étape ou modifier le système obtenu en remplaçant les pôles P_i extérieurs au cercle unité par $\frac{1}{\overline{P}_i}$, ce qui ne modifie pas le module de la réponse en fréquence à une constante près. Dans ce dernier cas il faut en général reprendre la procédure d'optimisation pour aboutir à l'optimum.

La minimisation de l'erreur quadratique moyenne peut être appliquée à d'autres fonctions que la réponse en fréquence, par exemple le temps de propagation de groupe [8].

7.3.2 Approximation au sens de Tchebycheff

Ce critère correspond à une limitation de l'amplitude des ondulations de la réponse en fréquence du filtre dans certaines plages de fréquence, ce qui est le cas le plus courant.

Une méthode élégante consiste à faire appel à l'algorithme utilisé pour le calcul des coefficients des filtres RIF à phase linéaire, l'algorithme de Remez.

La technique de calcul consiste à partir d'une fonction de filtrage initiale $H_0(Z)$ proche de la fonction H(Z) cherchée. Cette fonction peut être par exemple de

type elliptique et avoir été calculée par la méthode du paragraphe 7.2.4 en utilisant un gabarit adapté ; elle s'écrit :

$$H_0(Z) = \frac{N_0(Z)}{D_0(Z)}$$

Les zéros des fonctions de filtrage étant en général sur le cercle unité, le numérateur $N(Z)$ peut être considéré comme la fonction de transfert d'un filtre RIF à phase linéaire.

La première étape de la technique itérative consiste à calculer une nouvelle valeur du numérateur, $N_1(Z)$, par l'algorithme donné précédemment. Dans ce calcul $D_0(f)$ est la fonction à approcher en bande passante et $\frac{1}{|D_0(f)|}$ est utilisée comme pondération.

Ensuite on recherche une nouvelle valeur du dénominateur $D_1(Z)$. On peut chercher directement une fonction qui approche $|N_1(f)|$ dans la bande passante en utilisant le programme de calcul de filtre RIF. Une méthode plus satisfaisante consiste à utiliser une adaptation des techniques de calcul des filtres analogiques.

En supposant que la fonction cherchée $H(f)$, qui s'écrit :

$$H(f) = \frac{N(f)}{D(f)}$$

soit telle que :

$$|H(f)| \leqslant 1$$

on peut poser :

$$|G(f)|^2 = |D(f)|^2 - |N(f)|^2$$

et il vient :

$$|H(f)|^2 = \frac{1}{1 + \left|\frac{G(f)}{N(f)}\right|^2}$$

La figure 7.9 représente les fonctions $|G(f)|$ et $|N(f)|$ dans le cas d'un filtre passe-bas.

Les zéros de la fonction $G(Z)$ sont sur le cercle unité et, pour la calculer, on peut utiliser une procédure basée sur un programme de calcul de filtre RIF. La fonction de pondération est déterminée à partir de $\frac{1}{|N(f)|}$.

En optimisant ainsi alternativement en bande affaiblie et en bande passante, on aboutit, en quelques itérations à la fonction de filtrage cherchée.

Les coefficients du filtre sont ensuite obtenus en ne conservant pour $H(Z)$ que les pôles qui sont à l'intérieur du cercle unité, afin que le filtre soit stable.

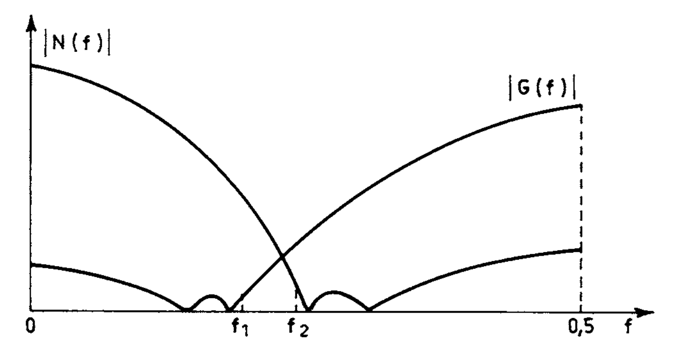


FIG. 7.9. *Fonctions* N(f) *et* G(f) *pour un filtre passe-bas*

La figure 7.10 montre un filtre de voie téléphonique qui a été calculé en utilisant cette méthode. Des techniques d'optimisation plus générales peuvent aussi conduire au filtre cherché, notamment la programmation linéaire [3].

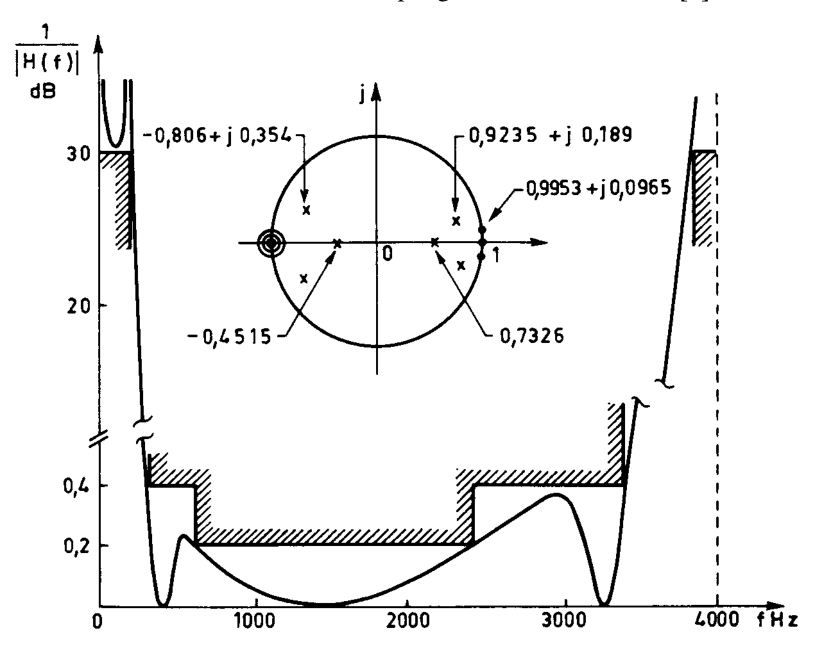


FIG. 7.10. *Filtre calculé par une technique itérative*

7.4 FILTRES BASÉS SUR LES FONCTIONS SPHÉROÏDALES

Au lieu de chercher à approcher un gabarit ou une fonction, le critère de calcul des coefficients peut être la maximisation de la concentration de l'énergie dans une bande de fréquence.

Soit λ un scalaire représentant cette concentration et défini par l'expression :

$$\lambda = \int_{-fc}^{fc} H(f)\,\overline{H(f)}\,df \Big/ \int_{-\frac{1}{2}}^{\frac{1}{2}} H(f)\,\overline{H(f)}\,df \qquad (7.38)$$

où $[-f_c, f_c]$ est la bande dans laquelle on cherche à concentrer l'énergie.

Pour :

$$H(f) = \sum_{n=-P}^{P} a_n\, e^{-j2\pi fn}$$

il vient, par calcul direct :

$$\lambda = \frac{\displaystyle\sum_{n=-P}^{P}\sum_{m=-P}^{P} a_n a_m \frac{\sin(n-m)2\pi f_c}{(n-m)\pi}}{\displaystyle\sum_{n=-P}^{P} a_n^2} \qquad (7.39)$$

Ou encore sous forme matricielle :

$$A^t R A = \lambda A^t A.$$

C'est une équation aux valeurs propres et les coefficients du filtre sont les éléments du vecteur propre associé à la plus grande valeur propre de la matrice carrée R, dont les éléments sont les termes $\dfrac{\sin(n-m)2\pi f_c}{(n-m)\pi}$.

Les éléments des vecteurs propres de la matrice R sont appelés séquences sphéroïdales [9].

Un filtre RIF a ainsi été obtenu. Il est également possible d'obtenir un filtre RII. En effet, soit le filtre purement récursif tel que :

$$|H(f)|^2 = \frac{1}{1 + \left|\displaystyle\sum_{n=1}^{N} b_n e^{j2\pi fn}\right|^2}$$

pour lequel les coefficients sont calculés de manière à minimiser l'énergie du dénominateur dans la bande $[-f_c, f_c]$, sous la condition $|H(f_c)|^2 = \dfrac{1}{2}$. Alors le calcul précédent peut être reconduit, en calculant les coefficients $b_n (1 \leqslant n \leqslant N)$ à partir des éléments du vecteur propre associé à la plus petite valeur propre de la matrice sphéroïdale. D'abord le facteur d'échelle du vecteur propre est choisi tel que : $|H(f_c)|^2 = \dfrac{1}{2}$. Ensuite on calcule les pôles de l'extension analytique de $|H(f)|^2$ et le filtre $H(Z)$ est obtenu en ne conservant que ceux qui sont à l'intérieur du cercle unité, pour fournir un filtre stable.

Les calculs peuvent être simplifiés par l'utilisation de procédures itératives et l'exploitation des propriétés de la matrice sphéroïdale [9].

Exemple :

Soit : $N = 4 ; \quad f_e = 1 ; \quad f_c = 0{,}1.$

Le vecteur propre minimal V_{min} s'écrit :

$$V_{min}^t = [1{,}0 \quad -2{,}773 \quad -2{,}773 \quad 1{,}0]$$

Si T désigne la matrice dont les éléments sont les termes $e^{j2\pi f_c(n-m)}$ avec $1 \leqslant n, m \leqslant N$, le facteur d'échelle correspondant à l'égalité $V_{min}^t T\, V_{min} = 1$ a pour valeur 10,46.

Après factorisation de l'extension analytique $H(Z)\, H(Z^{-1})$, la fonction de transfert du filtre obtenu est la suivante :

$$H(Z) = \frac{0{,}0704}{(Z - 0{,}73 + j0{,}446)\,(Z - 0{,}73 - j0{,}446)\,(Z - 0{,}741)}$$

La méthode de calcul, présentée pour le filtrage passe-bas, s'étend au cas des filtres passe-bande.

7.5 LES STRUCTURES REPRÉSENTANT LA FONCTION DE TRANSFERT

Les filtres RII peuvent être réalisés par des circuits qui effectuent directement les opérations représentées dans l'expression de leur fonction de transfert. Le terme Z^{-1} correspond à un retard d'une période d'échantillonnage, réalisé par une mise en mémoire ; les coefficients à mettre en œuvre dans les circuits sont ceux de la fonction de transfert, avec le même signe pour le numérateur et le signe opposé pour le dénominateur.

Seules les structures canoniques, c'est-à-dire celles qui demandent le minimum d'opérateurs élémentaires, circuits de calcul et mémoires, sont examinées.

7.5.1 Les structures directes

Elles correspondent à une réalisation globale de la fonction de transfert en Z. Soit à réaliser le filtre purement récursif de fonction de transfert :

$$H(Z) = \frac{1}{1 + \sum_{i=1}^{N} b_i Z^{-i}}$$

Un nombre de la suite de sortie $y(n)$ est obtenu à partir des nombres de la suite d'entrée $x(n)$ par la relation :

$$y(n) = x(n) - \sum_{i=1}^{N} b_i y(n-i)$$

qui donne les opérations à effectuer dans la réalisation du filtre. Le schéma correspondant est donné par la figure 7.11. Le circuit comprend N mémoires de données pour stocker les $y(n-i)$ ($i = 1, \ldots, N$). Le calcul de chaque élément de la suite de sortie nécessite N multiplications et N additions.

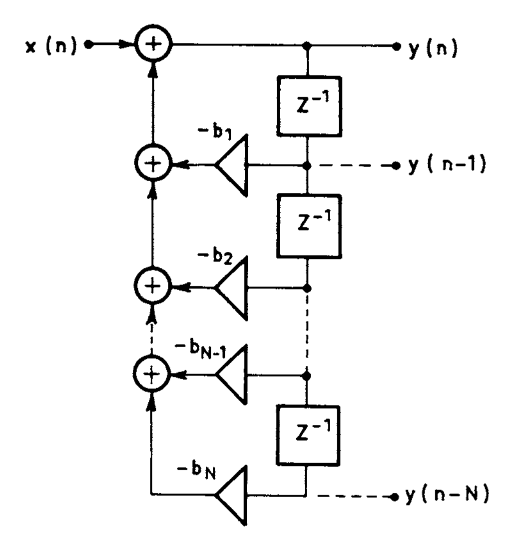


FIG. 7.11. *Circuit de filtre purement récursif*

Un filtre RII général peut être considéré comme la mise en cascade d'un filtre purement récursif et d'un filtre RIF. Quand le numérateur et le dénominateur de la fonction de transfert sont de même degré, le circuit obtenu est celui de la figure 7.12. Il correspond à la fonction de transfert en Z suivante :

$$H(Z) = \frac{\sum_{i=0}^{N} a_i Z^{-i}}{1 + \sum_{i=1}^{N} b_i Z^{-i}}$$

Comme le dénominateur est calculé en premier, la structure est dite D-N.

L'ordre des opérations peut être inversé et le numérateur calculé en premier. La structure dite N-D se déduit de la précédente par une transposition ; elle est donnée par la figure 7.13. Les nombres stockés dans les mémoires sont des sommes partielles. Une particularité intéressante est que chaque nombre $y(n)$ ou $x(n)$ est multiplié par tous les coefficients successivement, ce qui peut simplifier la mise en œuvre de la multiplication.

Comme les cellules du second ordre, ces structures peuvent être décrites par les équations d'état (4.34) et (4.37), en introduisant les variables $u_i(n)$ et $v_i(n)$ avec $1 \leqslant i \leqslant N$.

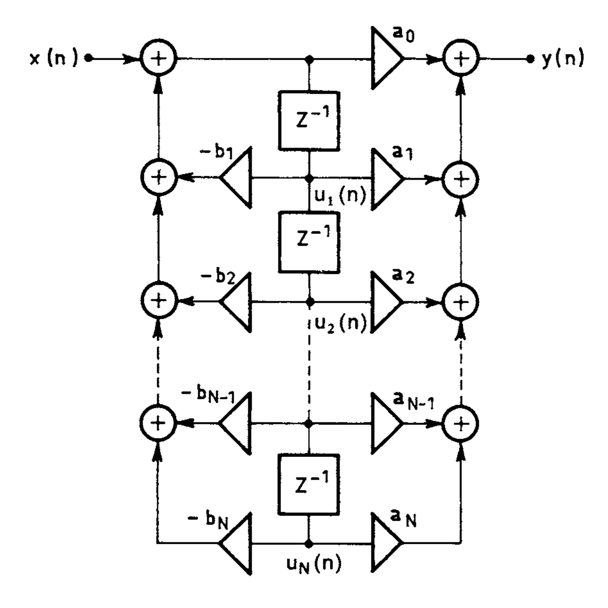


FIG. 7.12. *Filtres RII en structure directe D-N*

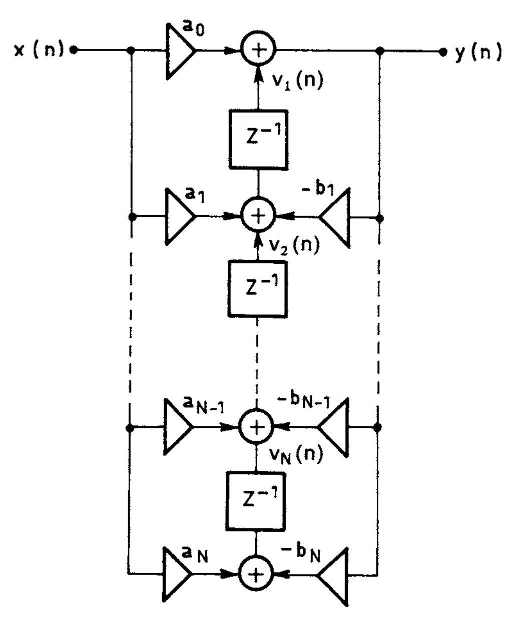


FIG. 7.13. *Filtre RII en structure directe N-D*

La matrice du système A s'écrit :

$$A = \begin{bmatrix} -b_1 & -b_2 & \ldots & -b_N \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix} \tag{7.40}$$

De même il vient :

$$B = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \quad d = a_0$$

$$C^t = (a_1 - a_0 b_1, a_2 - a_0 b_2, \ldots, a_N - a_0 b_N)$$

En pratique les structures directes sont peu utilisées car elles présentent des difficultés de réalisation, liées à la limitation du nombre de bits des coefficients, qui conduisent à préférer les structures décomposées.

7.5.2 Les structures décomposées

Au lieu de réaliser H(Z) directement, on peut effectuer une décomposition en somme ou produit de fonctions élémentaires du premier ou de second ordre réalisées séparément.

La décomposition en produit correspond à la structure cascade où le filtre est réalisé par une suite de cellules du premier et du second ordre :

$$H(Z) = a_0 \frac{\prod_{i=1}^{N} (1 - Z_i Z^{-1})}{\prod_{i=1}^{N} (1 - P_i Z^{-1})} = a_0 \ldots . \frac{1 - Z_i Z^{-1}}{1 - P_i Z^{-1}} \ldots \tag{7.41}$$

$$\ldots \frac{1 - 2Re(Z_j) Z^{-1} + |Z_j|^2 Z^{-2}}{1 - 2Re(P_j) Z^{-1} + |P_j|^2 Z^{-2}} \ldots$$

Cette structure est la plus utilisée car elle présente, en plus de sa modularité des caractéristiques avantageuses de faible sensibilité aux arrondis des coefficients et au bruit de calcul.

La fonction H(Z) se décompose aussi en fractions rationnelles; il vient :

$$H(Z) = a_0 + \ldots + \frac{\alpha_i}{1 - P_i Z^{-1}} + \ldots + \frac{\alpha_j + \beta_j Z^{-1}}{1 - 2Re(P_j) Z^{-1} + |P_j|^2 Z^{-2}} + \ldots \tag{7.42}$$

La réalisation correspond à la mise en parallèle de M cellules élémentaires comme indiqué sur la figure 7.14.

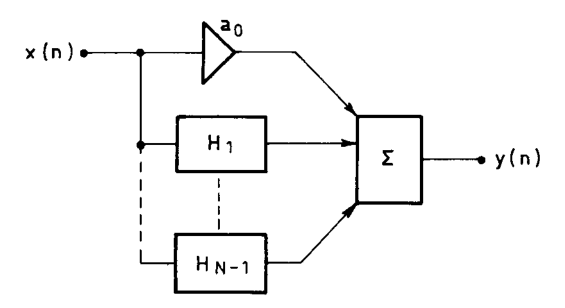


FIG. 7.14. *Structure parallèle*

Les nombres $y(n)$ sont obtenus par sommation des nombres isus des différentes cellules auxquelles sont appliqués les nombres d'entrée $x(n)$.

Le choix entre ces différentes formes de réalisation est conditionné par les facilités de mise en œuvre, par les incidences de la limitation du nombre de bits dans la représentation des coefficients sur les caractéristiques du filtre réalisé et par la puissance du bruit de calcul produit.

7.5.3 Structure à base de déphaseurs

Les fonctions de transfert de type Butterworth, Chebycheff ou elliptique peuvent se décomposer en une somme de deux déphaseurs [10]. Pour une telle fonction, il vient :

$$H(Z) = \frac{N(Z)}{D(Z)} = \frac{1}{2}\,[A_1(Z) + A_2(Z)] \qquad (7.43)$$

où $A_1(Z)$ et $A_2(Z)$ sont des fonctions de transfert de déphaseurs.

Le calcul de $A_1(Z)$ et $A_2(Z)$ à partir de $H(Z)$ fait intervenir la fonction complémentaire $G(Z) = \dfrac{M(Z)}{D(Z)}$ telle que :

$$|G(f)|^2 = 1 - |H(f)|^2 \qquad (7.44)$$

On suppose que la fonction de départ $H(Z)$ est telle que $N(Z)$ est un polynôme symétrique et $M(Z)$ est antisymétrique, c'est-à-dire :

$$\overline{N}(Z) = Z^N N(Z); \quad \overline{M}(Z) = -Z^N M(Z) \qquad (7.45)$$

Dans ces conditions, il vient, en combinant (VII-44) et (VII-45) :

$$\overline{N}(Z)N(Z) + \overline{M}(Z)M(Z) = \overline{D}(Z)D(Z) \qquad (7.46)$$

et

$$[N(Z) + M(Z)]\,[N(Z) - M(Z)] = Z^{-N} D(Z^{-1})\, D(Z) \qquad (7.47)$$

On peut remarquer que les zéros de $N(Z) + M(Z)$ et $N(Z) - M(Z)$ sont conjugués harmoniques, et que ce sont les zéros de $D(Z)$ et leurs inverses. En désignant par P_i ($i = 1, \ldots, N$) les pôles du filtre, donc les zéros de $D(Z)$, on peut écrire, à une constante près :

$$N(Z) + M(Z) = \prod_{i=1}^{r} (1 - Z^{-1}P_i) \prod_{i=r+1}^{N} (Z^{-1} - P_i) \qquad (7.48)$$

et

$$N(Z) - M(Z) = \prod_{i=1}^{r} (Z^{-1} - P_i) \prod_{i=r+1}^{N} (1 - Z^{-1}P_i)$$

où r est le nombre de zéros à l'intérieur du cercle unité pour le polynôme $N(Z) + M(Z)$. En divisant par $D(Z)$, on obtient :

$$H(Z) + G(Z) = \frac{\prod_{i=r+1}^{N} (Z^{-1} - P_i)}{\prod_{i=r+1}^{N} (1 - Z^{-1}P_i)} \qquad (7.49)$$

et, de même

$$H(Z) - G(Z) = \frac{\prod_{i=1}^{r} (Z^{-1} - P_i)}{\prod_{i=1}^{r} (1 - Z^{-1}P_i)} \qquad (7.50)$$

Les déphaseurs $A_1(Z)$ et $A_2(Z)$ ont les expressions suivantes :

$$A_1(Z) = \prod_{i=r+1}^{N} \frac{Z^{-1} - P_i}{1 - Z^{-1}P_i}\ ; \quad A_2(Z) = \prod_{i=1}^{r} \frac{Z^{-1} - P_i}{1 - Z^{-1}P_i} \qquad (7.51)$$

Finalement, le filtre $H(Z)$ et son complément $G(Z)$ sont obtenus par le schéma de la figure 7.15.

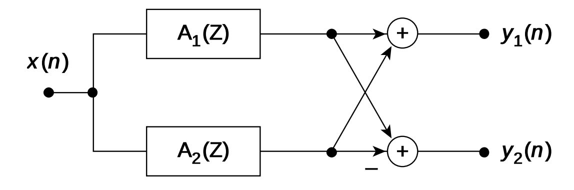


FIG. 7.15. *Réalisation d'un filtre RII et du filtre complémentaire par un couple de déphaseurs*

La procédure générale pour le calcul des déphaseurs à partir d'un filtre elliptique par exemple est la suivante :

– Calculer une fonction de transfert $H(Z) = \dfrac{N(Z)}{D(Z)}$ de filtre elliptique d'ordre N impair.

– Calculer les coefficients du polynôme antisymétrique $M(Z)$ à partir de $N(Z)$ et $D(Z)$ en utilisant (7.46).

– Déterminer les inverses des pôles de $H(Z)$ qui sont des racines du polynôme $N(Z) + M(Z)$

– Calculer $A_1(Z)$ et $A_2(Z)$ par la relation (7.51).

Une approche simplifiée, quand l'ordre N est peu élevé, consiste à rechercher directement $A_1(Z)$ et $A_2(Z)$ par combinaison des pôles. Ainsi, pour :

$$H(Z) = 0{,}0546 \; \frac{1 + 1{,}8601\,Z^{-1} + 2{,}9148\,Z^{-2} + 2{,}9148\,Z^{-3} + 1{,}8601\,Z^{-4} + Z^{-5}}{(1 - 0{,}4099\,Z^{-1})\,(1 - 0{,}6611\,Z^{-1} + 0{,}4555\,Z^{-2})\,(1 - 0{,}4993\,Z^{-1} + 0{,}8448\,Z^{-1})}$$

il vient :

$$A_1(Z) = \frac{0{,}4555 - 0{,}6611\,Z^{-1} + Z^{-2}}{1 - 0{,}6611\,Z^{-1} + 0{,}4555\,Z^{-2}}\;;$$

$$A_2(Z) = \frac{(-0{,}4099 + Z^{-1})\,(0{,}8448 - 0{,}4993\,Z^{-1} + Z^{-2})}{(1 - 0{,}4099\,Z^{-1})\,(1 - 0{,}4993\,Z^{-1} + 0{,}8448\,Z^{-2})}$$

La structure à base de déphaseurs est intéressante car elle fournit deux filtres complémentaires avec les mêmes calculs, ce qui est utile dans les bancs de filtres, comme indiqué aux chapitres 11 et 12. De plus, elle est moins sensible que les autres structures aux arrondis des coefficients.

Il est remarquable d'observer que les filtres décomposables en somme de déphaseurs sont entièrement définis par leurs pôles.

En fait, la réalisation en somme de déphaseurs de la figure (7.15) est la réalisation la plus efficace pour un filtre elliptique, puisqu'elle nécessite un nombre de multiplications égal à l'ordre du filtre.

7.6 LIMITATION DU NOMBRE DE BITS DES COEFFICIENTS

La mise en œuvre des opérations de filtrage implique la limitation du nombre de bits des coefficients du filtre qui constituent un des termes des multiplications. L'incidence sur la complexité est importante car la multiplication est souvent le facteur le plus critique. Il faut donc rechercher le nombre de bits minimal qui permette de satisfaire aux contraintes imposées à la fonction de filtrage.

La limitation du nombre de bits du facteur d'échelle a_0 se traduit par une modification du gain du filtre;, mais n'affecte pas la forme de la réponse en fréquence. Le gain du filtre étant spécifié avec une certaine tolérance à une fréquence

donnée, par exemple 800 Hz pour une voie téléphonique, il faut s'assurer que la représentation binaire de a_0 permet de satisfaire cette contrainte.

La limitation du nombre de bits des autres coefficients modifie la fonction de transfert en introduisant des polynômes parasites $e_N(Z)$ et $e_D(Z)$ au numérateur et au dénominateur. On a en fait la fonction de transfert $H_R(Z)$ telle que :

$$H_R(Z) = \frac{N(Z) + e_N(Z)}{D(Z) + e_D(Z)} \qquad (7.52)$$

Si l'on désigne par δa_i et δb_i $(1 \leqslant i \leqslant N)$ les erreurs d'arrondi faites sur les coefficients, ces fonctions parasites s'écrivent en fonction de la fréquence normalisée $(f_e = 1)$:

$$e_N(f) = \sum_{i=1}^{N} \delta a_i e^{-j2\pi fi}; \quad e_D(f) = \sum_{i=1}^{N} \delta b_i e^{-j2\pi fi}$$

En fait ces expressions constituent les développements en série de Fourier de fonctions périodiques de la fréquence. L'égalité de Bessel-Parseval (1.7) qui relie la puissance d'un signal à celle de ses composantes permet d'écrire, si $f_e = 1$:

$$\int_0^1 |e_N(f)|^2\, df = \sum_{i=1}^{N} |\delta a_i|^2$$

Si q désigne l'échelon de quantification :

$$|\delta a_i| \leqslant \frac{q}{2}$$

et une borne est obtenue pour $|e_N(f)|$ par :

$$|e_N(f)| \leqslant N\,\frac{q}{2} \qquad (7.53)$$

Une estimation statistique σ de $|e_N(f)|$ peut être obtenue en considérant les δa_i comme des variables aléatoires uniformément réparties sur l'intervalle $\left[-\frac{q}{2}, \frac{q}{2}\right]$. Elle est évaluée à partir de la valeur efficace de la fonction $e_N(f)$. Il vient :

$$\sigma^2 = \int_0^1 |e_N(f)|^2\, df = N\,\frac{q^2}{12}$$

D'où :

$$\sigma = \frac{q}{2}\sqrt{\frac{N}{3}} \qquad (7.54)$$

Cette estimation est valable à la fois pour $|e_N(f)|$ et $|e_D(f)|$; elle est nettement inférieure à la borne (7.53) donnée ci-dessus et en fait beaucoup plus proche de la réalité, dès que N dépasse quelques unités.

Les conséquences de l'arrondi des coefficients peuvent être analysées séparément pour le numérateur et le dénominateur de la fonction de transfert, en considérant la bande affaiblie d'une part et la bande passante de l'autre. En effet en examinant la configuration des pôles et des zéros dans le plan des Z, on observe

que les pôles déterminent la réponse du filtre en bande passante et les zéros en bande affaiblie.

– En bande affaiblie, l'arrondi des coefficients du dénominateur peut être négligé et, avec comme variable $\omega = 2\pi f$, il vient :

$$H_R(\omega) = \frac{N(\omega) + e_N(\omega)}{|D(\omega)|}$$

L'erreur sur la réponse est alors estimée par :

$$|H_R(\omega) - H(\omega)| \simeq \frac{\sigma}{|D(\omega)|}$$

Si le gabarit impose que les ondulations en bande affaiblie soient inférieures en module à δ_2, en partageant la tolérance en deux parties égales, l'une pour les ondulations en l'absence d'erreur d'arrondi sur les coefficients et l'autre pour tenir compte de l'erreur due à cet arrondi, il vient :

$$\frac{\sigma}{D(\omega)} < \frac{\delta_2}{2} \qquad (7.55)$$

– En bande passante, l'arrondi des coefficients du numérateur peut être négligé :

$$H_R(\omega) \simeq \frac{N(\omega)}{D(\omega) + e_D(\omega)} \simeq \frac{N(\omega)}{D(\omega)} \left[1 - \frac{e_D(\omega)}{D(\omega)}\right] \qquad (7.56)$$

Si les ondulations en bande passante doivent être inférieures en module à δ_1, l'inégalité suivante doit être vérifiée, en partageant encore en deux parties égales la tolérance :

$$\frac{\sigma}{|D(\omega)|} < \frac{\delta_1}{2} \qquad (7.57)$$

Cette condition est en général beaucoup plus contraignante que la précédente car la fonction $|D(\omega)|$ prend en bande passante des valeurs très faibles, d'autant plus que le filtre est plus sélectif. De plus, quand la bande passante est étroite, les coefficients peuvent prendre des valeurs importantes. Pour un passe-bas comme celui de la figure 7.16, on peut écrire :

$$D(Z) \simeq [1 - Z^{-1}]^N$$

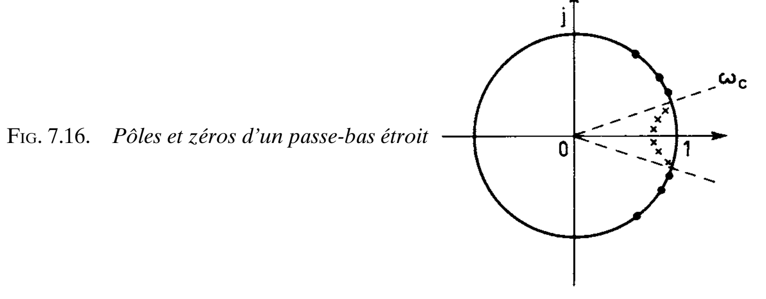


FIG. 7.16. *Pôles et zéros d'un passe-bas étroit*

et par suite :

$$b_i \simeq \frac{N!}{i!(N-i)!} \tag{7.58}$$

Dans ces conditions il faut un très grand nombre de bits pour pouvoir à la fois représenter les grandes valeurs des coefficients et faire une erreur de quantification très faible. C'est pourquoi on utilise presque exclusivement les structures décomposées faisant appel à des cellules du premier ou du second ordre.

Considérons d'abord la structure cascade correspondant à la décomposition (7.41) de la fonction de transfert qui, si l'ordre N du filtre est pair, s'écrit :

$$H(\omega) = \frac{N(\omega)}{D(\omega)} = \prod_{i=1}^{\frac{N}{2}} \frac{N_i(\omega)}{D_i(\omega)}$$

et où les polynômes $N_i(\omega)$ et $D_i(\omega)$ sont du second degré.

En bande passante, il vient, si l'arrondi des coefficients des polynômes $N_i(\omega)$ est négligé :

$$H_R(\omega) \simeq \prod_{i=1}^{\frac{N}{2}} \frac{N_i(\omega)}{D_i(\omega) + e_i(\omega)}$$

ou encore :

$$H_R(\omega) \simeq \frac{N(\omega)}{D(\omega)} \left[1 - \sum_{i=1}^{\frac{N}{2}} \frac{e_i(\omega)}{D_i(\omega)}\right] \tag{7.59}$$

Or, d'après (7.53) on a :

$$|e_i(\omega)| \leqslant q$$

et l'erreur relative globale $e(\omega)$ sur la réponse en fréquence se trouve bornée par :

$$|e(\omega)| \leqslant q \,.\, \sum_{i=1}^{\frac{N}{2}} \frac{1}{|D_i(\omega)|} \tag{7.60}$$

Cette expression met bien en évidence l'avantage de la structure décomposée puisque la borne de l'erreur est proportionnelle à :

$$\sum_{i=1}^{\frac{N}{2}} \frac{1}{|D_i(\omega)|}$$

au lieu d'être, d'après (7.56), proportionnelle à :

$$\prod_{i=1}^{\frac{N}{2}} \frac{1}{|D_i(\omega)|}$$

De plus la valeur absolue des coefficients du dénominateur ne peut dépasser 2 pour que le filtre soit stable.

Avec la structure parallèle, également en bande passante, on a :

$$H(\omega) \simeq \sum_{i=1}^{\frac{N}{2}} \frac{N'_i(\omega)}{D_i(\omega) + e_i(\omega)} \simeq \sum_{i=1}^{\frac{N}{2}} \frac{N'_i(\omega)}{D_i(\omega)}\left[1 - \frac{e_i(\omega)}{D_i(\omega)}\right]$$

Compte tenu du fait que les termes $\left|\frac{N'_i(\omega)}{D_i(\omega)}\right|$ ne sont pas très différents de l'unité, l'erreur faite dans ce cas est voisine de celle qui est faite avec la structure cascade.

En bande affaiblie, on obtient pour la structure cascade, en dehors des fréquences d'affaiblissement infini :

$$H_R(\omega) \simeq \prod_{i=1}^{\frac{N}{2}} \frac{N_i(\omega) + e_i(\omega)}{D_i(\omega)} \simeq \frac{N(\omega)}{D(\omega)}\left[1 + \sum_{i=1}^{\frac{N}{2}} \frac{e_i(\omega)}{N_i(\omega)}\right] \qquad (7.61)$$

Et pour la structure parallèle on peut faire l'estimation :

$$H_R(\omega) \simeq \sum_{i=1}^{\frac{N}{2}} \frac{N'_i(\omega) + e_i(\omega)}{D_i(\omega)} = \sum_{i=1}^{\frac{N}{2}} \frac{N'_i(\omega)}{D_i(\omega)} + \sum_{i=1}^{\frac{N}{2}} \frac{e_i(\omega)}{D_i(\omega)}$$

ou encore

$$H_R(\omega) \simeq \frac{N(\omega)}{D(\omega)}\left[1 + \sum_{i=1}^{\frac{N}{2}} \frac{e_i(\omega)}{N_i(\omega)} \prod_{\substack{j=1 \\ i \neq 1}}^{\frac{N}{2}} \frac{D_j(\omega)}{N_j(\omega)}\right] \qquad (7.62)$$

En faisant la comparaison entre (7.61) et (7.62), on remarque que les termes $\alpha_i = \prod_{\substack{i \neq 1 \\ j=1}}^{\frac{N}{2}} \frac{D_j(\omega)}{N_j(\omega)}$ peuvent prendre, en bande affaiblie, des valeurs beaucoup plus grandes que l'unité, par exemple au voisinage des zéros du filtre ; il en résulte que la structure parallèle est plus sensible aux erreurs d'arrondis que la structure cascade dans la bande affaiblie.

Finalement la structure cascade est celle qui permet de représenter les coefficients des filtres RII avec le nombre de bits le plus faible, c'est la plus utilisée.

7.7 NOMBRE DE BITS DES COEFFICIENTS EN STRUCTURE CASCADE

Pour faire apparaître des formules simples donnant le nombre de bits nécessaire à la représentation des coefficients d'un filtre passe-bas réalisé en structure cascade, il faut d'abord rappeler un certain nombre d'observations faites précédemment.

D'abord la fonction de transfert H(Z) d'un filtre elliptique s'écrit :

$$\mathrm{H}(\mathrm{Z}) = \prod_{i=1}^{\frac{\mathrm{N}}{2}} a_0^i \, \frac{1 + a_1^i \mathrm{Z}^{-1} + \mathrm{Z}^{-2}}{1 + b_1^i \mathrm{Z}^{-1} + b_2^i \mathrm{Z}^{-2}} \tag{7.63}$$

D'après les résultats du paragraphe VI.5, l'arrondi des coefficients a_1^i amène seulement un déplacement sur l'axe des fréquences de la pointe d'affaiblissement infini correspondante du filtre. Dans le calcul il est possible d'en tenir compte et ainsi l'effet de l'arrondi des coefficients du numérateur de la fonction de transfert en Z se trouve minimisé. Par suite c'est généralement le dénominateur de la fonction H(Z) qui impose le nombre de bits des coefficients. Les pôles de cette fonction doivent rester à l'intérieur du cercle unité dans le plan complexe et, comme indiqué au paragraphe 6.5, la limitation à b_c bits par arrondi correspond pour les coefficients à une quantification avec un échelon q tel que :

$$q = 2^{2 - b_c} \tag{7.64}$$

Le nombre de bits b_c inclut le signe. L'erreur qui résulte de cet arrondi sur la réponse en fréquence est bornée par la relation (7.60). En général, comme indiqué au début de ce paragraphe, l'erreur $e(\omega)$ est nettement inférieure à cette borne et pour obtenir une estimation réaliste il faut faire une évaluation statistique. Par analogie avec la distribution gaussienne où le rapport entre la valeur de crête et l'écart type est voisin de 4, on peut écrire :

$$|e(\omega)| \simeq \frac{q}{4} \sum_{i=1}^{\frac{\mathrm{N}}{2}} \frac{1}{|\mathrm{D}_i(\omega)|} \tag{7.65}$$

Les valeurs les plus faibles pour les termes $|\mathrm{D}_i(\omega)|$ sont atteintes en bande passante où la limite à considérer est δ_1. Si δ_{10} désigne l'amplitude des ondulations du filtre avant limitation du nombre de bits des coefficients il faut que :

$$|e(\omega)| \leqslant \delta_1 - \delta_{10}$$

et par suite, compte tenu de (7.64) et (7.65) il faut choisir b_c tel que :

$$b_c \simeq \log 2\left(\frac{1}{\delta_1 - \delta_{10}}\right) + \log 2\left(\operatorname*{Max}_{0 \leqslant \omega \leqslant \pi} \sum_{i=1}^{\frac{\mathrm{N}}{2}} \frac{1}{|\mathrm{D}_i(\omega)|}\right) \tag{7.66}$$

Parmi les $\frac{\mathrm{N}}{2}$ cellules du second ordre, la plus importante pour l'estimation considérée est celle qui a les pôles les plus proches du cercle unité et dont le gain à la résonance prend la valeur la plus élevée. Soient r et θ les coordonnées polaires des pôles de cette cellule, d'après la relation (6.29) :

$$\mathrm{H}_m = \max_{0 \leqslant \omega \leqslant \pi} \frac{1}{|\mathrm{D}_i(\omega)|} = \frac{1}{(1-r)} \cdot \frac{1}{(1+r)\sin\theta}$$

D'autre part, la largeur de bande à 3 décibels B_3 de cette cellule est donnée par la relation (6.30) :

$$B_3 = \frac{1-r}{\pi} \cdot f_e$$

Or il existe, dans les filtres à grande sélectivité en particulier, une relation directe entre la largeur de bande B_3 de cette cellule et la bande de transition du filtre Δf; l'approximation suivante peut être faite :

$$\Delta f \simeq 3B_3$$

et par suite :

$$1 - r \simeq \frac{\Delta f}{f_e} \tag{7.67}$$

Compte tenu du fait qu'il est généralement possible d'approcher θ par $2\pi f_1/f_e$, il vient :

$$H_m \simeq \frac{f_e}{\Delta f} \cdot \frac{1}{2\sin(2\pi f_1/f_e)} \tag{7.68}$$

Pour les autres cellules l'amplitude à la résonance est nettement plus faible et on peut raisonnablement faire l'approximation supplémentaire suivante :

$$\operatorname*{Max}_{0 \leqslant \omega \leqslant \pi} \sum_{i=1}^{\frac{N}{2}} \frac{1}{|D_i(\omega)|} \simeq \frac{f_e}{\Delta f} \cdot \frac{1}{2\sin\left(2\pi \frac{f_1}{f_e}\right)}$$

Et finalement :

$$b_c \simeq \log 2\left(\frac{1}{\delta_1 - \delta_{10}}\right) + \log 2\left[\frac{f_e}{\Delta f} \cdot \frac{1}{2\sin\left(2\pi \frac{f_1}{f_e}\right)}\right] \tag{7.69}$$

Si la tolérance δ_1 est partagée entre les ondulations du filtre avant arrondi des coefficients et les ondulations supplémentaires dues à cet arrondi, $\delta_{10} = \frac{\delta_1}{2}$, il vient :

$$b_c \simeq \log 2\left(\frac{1}{\delta_1}\right) + \log 2\left(\frac{f_e}{\Delta f}\right) + \log 2\left(\frac{1}{\sin\left(2\pi \frac{f_1}{f_e}\right)}\right) \tag{7.70}$$

Cette expression est à rapprocher de (5.46) pour les filtres RIF. La bande de transition normalisée contribue à augmenter b_c et les filtres RII demandent en général un plus grand nombre de bits pour la représentation des coefficients que les filtres RIF. De plus ce nombre de bits croît quand la bande passante diminue.

Les estimations ci-dessus ont été faites avec l'hypothèse d'une limitation par arrondi. Pour la réponse en fréquence $H_R(\omega)$ d'un filtre, ce qui importe le plus c'est le comportement des polynômes parasites $e_i(\omega)$ au voisinage immédiat des

fréquences qui minimisent les $D_i(\omega)$, comme le montre la relation (7.59). En pratique il se peut que l'on trouve des configurations de coefficients quantifiés qui donnent des résultats meilleurs que l'arrondi. Ces configurations qui peuvent conduire à des valeurs de b_c inférieures de plusieurs unités à l'estimation (7.69), peuvent être atteintes par exemple par une recherche systématique au voisinage de l'arrondi.

7.8 BRUIT DE CALCUL

Dans la réalisation des filtres RII, une autre limitation intervient, celle qui porte sur la capacité des mémoires de données et qui est à l'origine du bruit de calcul. L'analyse correspondante va être faite dans le cas de la structure D-N, mais les mêmes raisonnements s'appliquent à la structure N-D. Bien que cette structure possède des avantages spécifiques, liés principalement au calcul des sommes partielles et à l'enchaînement des multiplications, c'est la structure D-N qui est la plus utilisée car elle est en général plus facile à concevoir, à mettre en œuvre et à vérifier.

En présence d'un signal, c'est-à-dire pour des valeurs non nulles de $x(n)$, l'opération d'arrondi avant mise en mémoire avec le pas de quantification q, est équivalente à la superposition au signal d'entrée d'un signal d'erreur $e(n)$ tel que $|e(n)| \leqslant \frac{q}{2}$, supposé à spectre uniforme et de puissance : $\sigma^2 = \frac{q^2}{12}$.

Si d'autres arrondis interviennent, par exemple dans les multiplications, il apparaît clairement que les signaux d'erreur produits sont à ajouter soit au signal d'entrée, soit au signal de sortie suivant qu'ils correspondent aux coefficients de la partie récursive ou non récursive, comme le montre la figure 7.17. Par suite pour simplifier l'analyse, seul le cas de la quantification unique est considéré, étant donné qu'il est toujours possible de se ramener facilement à ce cas en modifiant la puissance du bruit injecté.

Le signal d'erreur appliqué à l'entrée du filtre subit la fonction de filtrage et la puissance du bruit de calcul en sortie s'écrit, en appliquant la relation (4.25) :

$$B_c = \frac{q^2}{12}\int_0^1 \left|\frac{N(f)}{D(f)}\right|^2 df \qquad (7.71)$$

ou encore en fonction de la suite $h(k)$, réponse impulsionnelle du filtre :

$$B_c = \frac{q^2}{12}\sum_{k=0}^{\infty} |h(k)|^2 \qquad (7.72)$$

La réalisation, en structure cascade offre un certain nombre de possibilités pour réduire cette puissance de bruit [11].

Quand le filtre est réalisé par mise en cascade de $\frac{N}{2}$ cellules du second ordre, le bruit de calcul produit dans une cellule subit la fonction de filtrage de cette cel-

lule et des suivantes. Dans ce cas il faut remarquer que l'amplitude, ou niveau, à l'entrée de chaque cellule varie avec le rang de la cellule et la fréquence du signal considéré.

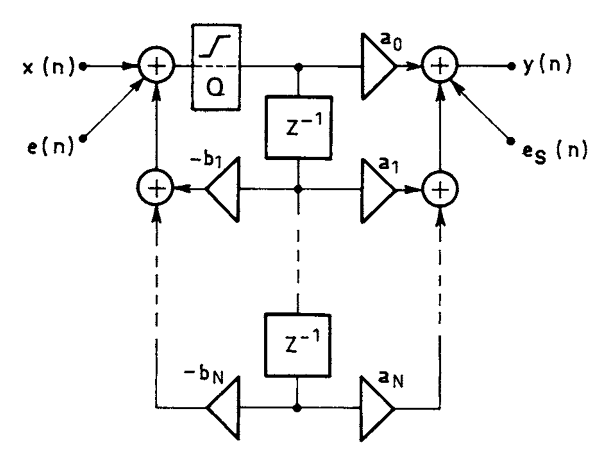


FIG. 7.17. *Erreurs d'arrondi dans la structure D-N*

La figure 7.18 donne le diagramme des amplitudes à une fréquence donnée dans la bande passante telle que l'amplitude prend la valeur 1 à l'entrée et à la sortie du filtre. Si la procédure d'arrondi est la même pour toutes les cellules le bruit produit est le même et les contributions s'ajoutent. Le bruit total en sortie du filtre dans ces conditions a pour puissance :

$$B_c = \frac{q^2}{12} \sum_{i=1}^{\frac{N}{2}} \left(\int_0^1 \prod_{i=j}^{\frac{N}{2}} \left| \frac{N_i(f)}{D_i(f)} \right|^2 df \right) \tag{7.73}$$

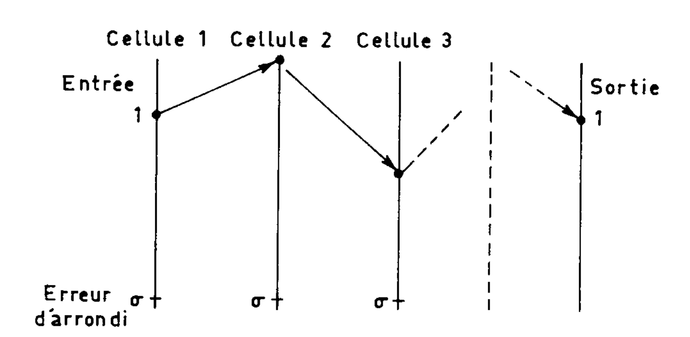


FIG. 7.18. *Diagramme des niveaux à une fréquence de la bande passante*

Il est important de constituer la cascade des cellules de façon à minimiser le bruit de calcul total.

On peut agir sur 3 paramètres :

– l'appairage des pôles et des zéros pour constituer une cellule,
– l'ordre dans lequel sont rangées les cellules,
– le facteur d'échelle affecté à chaque cellule.

Ces trois paramètres vont être examinés successivement :

– Appairage des pôles et zéros; il faut minimiser l'ensemble des produits :

$$P_j(f) = \prod_{i=j}^{\frac{N}{2}} \left| \frac{N_i(f)}{D_i(f)} \right|^2$$

ce qui conduit à minimiser chacun des facteurs, et en particulier obtenir la plus faible valeur maximale pour chaque facteur. Cette condition est approximativement remplie par la procédure très simple qui consiste à associer au pôle le plus proche du cercle unité le zéro le plus voisin, au pôle suivant le zéro restant le plus voisin et ainsi de suite, comme le montre la figure 7.19.

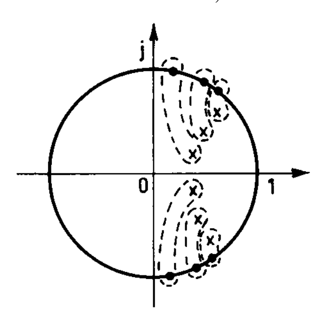


FIG. 7.19. *Appairage des pôles et zéros dans la structure cascade*

– Détermination de l'ordre des cellules : le facteur qui a la plus forte contribution au bruit total est souvent celui qui a la valeur maximale la plus élevée : il peut être intéressant de le mettre en début de chaîne pour qu'il ne contribue qu'une seule fois dans la somme totale suivant l'expression (7.73) et d'interconnecter les cellules par ordre de maximum décroissant.

Il faut remarquer cependant que la relation (7.73) suppose que l'arrondi est fait avec le même échelon pour toutes les cellules, ce qui n'est plus vérifié dans la réalité avec les recadrages des nombres dans les mémoires internes qu'entraîne l'introduction des facteurs d'échelle, pour éviter l'écrêtage des signaux dans chaque cellule. En toute rigueur la détermination de l'ordre des cellules nécessite la connaissance de ces facteurs d'échelle et l'évaluation de la puissance du bruit de calcul total dans chaque cas particulier. On imagine la complexité du problème pour les filtres d'ordre élevé. Aussi l'hypothèse de l'interconnexion des cellules par ordre maximum décroissant est conservée ici, car, d'une part elle conduit à des évaluations simples et généralement réalistes, d'autre part elle entraîne principalement la superposition d'un bruit ayant une répartition spectrale constante au signal présent à l'entrée du filtre, ce qui peut être intéressant en pratique, surtout pour le filtrage à bande étroite.

Finalement ces deux règles très simples d'appairage et d'interconnexion des cellules constituent une approximation suffisante pour la réalisation.

– Calcul des facteurs d'échelle : ce sont les paramètres qui commandent le cadrage des nombres dans les mémoires de données internes. C'est le dernier élément nécessaire pour évaluer le bruit de calcul et déterminer la capacité des mémoires en fonction des spécifications.

La limitation de la capacité des mémoires de données fait que l'amplitude du signal à l'intérieur du filtre est limitée à une valeur A_m, qui est une amplitude d'écrêtage, c'est-à-dire que tout nombre supérieur à A_m avant la mise en mémoire est ramené à cette valeur pour être mis en mémoire. Il résulte de cette opération une distorsion harmonique du signal qui souvent n'est pas tolérable et qu'il faut chercher à éviter.

Le signal appliqué à la première cellule doit avoir une amplitude telle que multipliée par le gain de cette cellule, elle ne conduise pas à un écrêtage inadmissible du signal; par suite il faut multiplier le signal d'entrée par un facteur préliminaire a_0^0. La fonction de transfert du filtre d'ordre N réalisé en $\frac{N}{2}$ cellules du 2e ordre peut s'écrire, pour un filtre elliptique, sous la forme :

$$H(\omega) = a_0^0 \prod_{i=1}^{\frac{N}{2}} a_0^i \frac{N_i(\omega)}{D_i(\omega)}$$

avec :

$$\frac{N_i(\omega)}{D_i(\omega)} = \frac{1 + a_1^i e^{-j\omega} + e^{-2j\omega}}{1 + b_1^i e^{-j\omega} + b_2^i e^{-2j\omega}}$$

Étant donnée l'hypothèse d'un arrondi unique dans chaque cellule faite précédemment, la relation (7.73) donnant le bruit de calcul total devient :

$$B_c = \frac{q^2}{12} (a_0^0)^2 \sum_{i=1}^{\frac{N}{2}} \int_0^1 \prod_{i=j}^{\frac{N}{2}} (a_0^i)^2 \left| \frac{N_i(f)}{D_i(f)} \right|^2 df \qquad (7.74)$$

Avec la structure D-N, le schéma du filtre est celui de la figure 7.20, où sont également représentés les points d'introduction des différentes puissances de bruit de calcul e_i avec $0 \leqslant i \leqslant \frac{N}{2}$. Les termes $a_0^i \left(0 \leqslant i \leqslant \frac{N}{2}\right)$ sont les facteurs d'échelle qui doivent être calculés de manière à éviter l'écrêtage du signal dans les cellules.

Pour calculer le facteur préliminaire a_0^0 il faut considérer la réponse de la première cellule $\frac{1}{D_1(\omega)}$ et tenir compte du fait que les nombres d'entrée du filtre $x(n)$ sont représentés par un nombre de bits limité et ont une amplitude supposée elle-même limitée à A_m, c'est-à-dire que :

$$|x(n)| \leqslant A_m$$

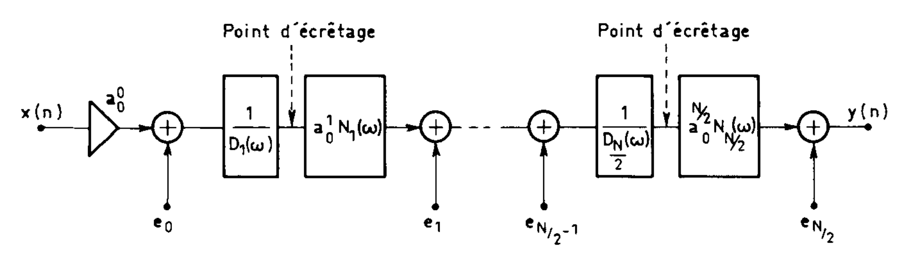


FIG. 7.20. *Schéma de structure cascade*

Si $y_1(n)$ désigne les nombres en sortie de la cellule de fonction de transfert $\dfrac{1}{D_1(\omega)}$ et si la suite $h(k)$ est la réponse impulsionnelle, on peut écrire :

$$|y_1(n)| \leqslant \sum_{k=0}^{\infty} |h(k)|\,|a_0^0 x(n-k)| \leqslant a_0^0\,A_m \sum_{k=0}^{\infty} |h(k)|$$

D'après les résultats donnés au paragraphe VI.2, pour une cellule à pôles complexes de coordonnées polaires (r, θ) il vient (VI.37) :

$$\sum_{k=0}^{\infty} |h(k)| \leqslant \frac{1}{(1-r)\sin\theta}$$

En prenant

$$a_0^0 = (1-r)\sin\ \theta \tag{7.75}$$

on garantit l'inégalité : $|y(n)| \leqslant A_m$.

On peut faire un choix moins restrictif pour le facteur a_0^0 en réduisant la condition d'absence d'écrêtage aux signaux sinusoïdaux ; dans ce cas c'est la valeur :

$$H_m^1 = \text{Max}\left|\frac{1}{D_1(\omega)}\right| = \frac{1}{(1-r^2)\sin\theta}$$

qu'il faut faire intervenir et dans ces conditions :

$$a_0^0 = (1-r^2)\sin\ \theta \tag{7.76}$$

Cette valeur est inférieure au double de la précédente.

Une troisième stratégie de cadrage est souvent utilisée, celle qui porte sur l'énergie des signaux [12].

D'après les résultats du paragraphe IV.3, l'énergie du signal $y_1(n)$ s'écrit :

$$\sum_{n=-\infty}^{\infty} y_1^2(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H_1(e^{j\omega})\,H_1(e^{-j\omega})\,(a_0^0)^2\,X(e^{j\omega})\,X(e^{-j\omega})\,d\omega$$

$$\leqslant \left[\sum_{k=-\infty}^{\infty} x^2(k)\right].(a_0^0)^2 \cdot \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{d\omega}{|D_1(\omega)|^2}$$

La condition :

$$(a_0^0)^2 \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{d\omega}{|D_1(\omega)|^2} = 1 \qquad (7.77)$$

conduit à l'expression du facteur a_0^0 suivante, d'après la relation (VI.36) :

$$(a_0^0)^2 = \frac{(1 - r^2)\,(1 + r^4 - 2r^2 \cos 2\,\theta)}{(1 + r^2)}$$

Exemple :

Soit $r = 0{,}845$; $\cos\theta = 0{,}396$ (fig. 7.6)

$$(1 - r) \sin\theta = 0{,}142$$
$$(1 - r^2) \sin\theta = 0{,}262$$
$$(3^{e}\ \text{cas})\ a_0^0 = 0{,}644$$

De ces trois stratégies de cadrage, qui correspondent aux normes des signaux L_1, L_∞ et L_2, la seconde constitue un bon compromis ; c'est celle qui est retenue dans la suite de ce paragraphe. Il faut aussi remarquer qu'elle facilite les vérifications expérimentales et les mises au point à l'aide de signaux sinusoïdaux.

On désigne par H_m^i et K_m^i les valeurs suivantes pour la cellule de rang i :

$$H_m^i = \max \left| \frac{1}{D_i(\omega)} \right| ;$$

$$K_m^i = \max \left| \frac{a_0^i N_i(\omega)}{D_i(\omega)} \right| ; \quad \text{avec} \quad 1 \leqslant i \leqslant \frac{N}{2}$$

Avec la technique de cadrage choisie :

$$a_0^0 = \frac{1}{H_m^1}$$

et pour les cellules à pôles complexes, d'après (VI.28) :

$$a_0^0 = (1 - b_2^1) \sqrt{1 - \frac{(b_1^1)^2}{4b_2^1}}$$

Il n'y a pas d'écrêtage en sinusoïdal dans la seconde cellule si l'inégalité suivante est vérifiée :

$$\frac{1}{H_m^1} \times K_m^1 \times H_m^2 \leqslant 1$$

Ce qui amène à prendre :

$$K_m^1 = \frac{H_m^1}{H_m^2}$$

Les termes K_m^i peuvent se calculer directement par (VI.40). Une valeur approchée peut être obtenue à l'aide de (VI.27) par :

$$K_m^1 \simeq a_0^1 . \left[a_1^1 - \frac{b_1^1 (1 + b_2^1)}{2 b_2^1} \right] . H_m^1$$

il s'en suit que :

$$a_0^1 \simeq \frac{(1 - b_2^2) \sqrt{1 - \dfrac{(b_1^2)^2}{4 b_2^2}}}{a_1^1 - \dfrac{b_1^1 (1 + b_2^1)}{2 b_2^1}}$$

Une estimation plus simple pour a_0^1 peut être obtenue en considérant l'inégalité :

$$\left| \frac{1}{H_m^1} \frac{1}{D_1(\omega)} N_1(\omega) \frac{1}{D_2(\omega)} \right| \leqslant \left| \frac{N_1(\omega)}{D_2(\omega)} \right|$$

Pour les fréquences voisines de celle qui minimise $D_2(\omega)$ on a souvent $|N_1(\omega) < 1|$, ce qui est facile à vérifier par la position des pôles et des zéros. Alors on peut prendre :

$$a_0^1 = \frac{1}{H_m^2}$$

Les facteurs d'échelle des cellules suivantes se déterminent de la même manière et finalement la procédure se résume comme suit :

1. Appliquer aux nombres d'entrée le facteur préliminaire $a_0^0 = \dfrac{1}{H_m^1}$.

2. Affecter à la cellule de rang $i \left(1 \leqslant i \leqslant \dfrac{N}{2} - 1 \right)$ le facteur d'échelle a_0^i tel que :

$$K_m^i = \frac{H_m^i}{H_m^{i+1}}$$

On peut souvent faire l'approximation :

$$a_0^i = \frac{1}{H_m^{i+1}} \tag{7.78}$$

Avec cette approximation, il faut cependant vérifier que les valeurs obtenues pour les derniers facteurs d'échelle ne sont pas trop faibles, afin de ne pas augmenter inutilement le bruit de calcul.

3. Calculer le facteur d'échelle $a_0^{\frac{N}{2}}$ de la dernière cellule pour que le filtre ait exactement le gain 1 en une fréquence de référence dans la bande passante.

En pratique, parmi les $\left(\dfrac{N}{2} + 1 \right)$ valeurs de a_0^i, $\dfrac{N}{2}$ sont prises comme puissances de deux, pour que la multiplication se ramène à un simple décalage.

La procédure simple exposée ci-dessus est suffisante dans de nombreux cas pratiques. Une méthode plus précise consiste à tenir compte de toutes les cellules précédentes pour déterminer un facteur d'échelle. Finalement il apparaît que le facteur d'échelle de la cellule de rang i est fonction des cellules qui la précèdent alors que les cellules qui suivent cette cellule filtrent le bruit qu'elle produit.

Une fois connus les facteurs d'échelle, tous les éléments qui interviennent dans la réalisation sont disponibles et la puissance de bruit de calcul en sortie du filtre peut être déterminée pour chaque valeur du nombre de bits des mémoires de données.

7.9 DÉTERMINATION DE LA CAPACITÉ DES MÉMOIRES INTERNES

A l'entrée du filtre le signal est constitué de nombres $x(n)$ qui sont représentés par un nombre de bits limité et égal à b_d. Par suite ce signal peut être considéré comme comprenant le signal idéal auquel se superpose un bruit de puissance B_1. Les nombres $x(n)$ sont supposés prendre des valeurs positives et négatives et limitées en valeur absolue à 1, c'est-à-dire :

$$|x(n)| \leqslant 1$$

La représentation du signal à b_d bits correspond à un échelon de quantification q tel que :

$$q = 2^{-b_d+1}$$

La puissance de bruit B_1 superposée au signal d'entrée est supposée égale à $k_0 (k_0 \geqslant 1)$ fois la puissance du bruit engendré par la quantification avec cet échelon, c'est-à-dire :

$$B_1 = k_0 \frac{2^{-2b_d}}{3} \tag{7.79}$$

En général le bruit de calcul, qui est le bruit ajouté par le filtre est à comparer au bruit déjà présent à l'entrée et ce qui importe c'est la dégradation du rapport signal à bruit lors de la traversée du filtre.

Comme expliqué précédemment, quelle que soit la précision des opérations arithmétiques, une opération d'arrondi doit intervenir sur les données internes avant la mise en mémoire; elle se traduit par la superposition à la suite représentant le signal, d'une erreur $e(n)$ et par conséquent par une dégradation du rapport signal à bruit à la traversée du filtre.

Dans un filtre en structure cascade, le bruit engendré par la limitation à b_i bits des données internes est supposé appliqué à l'entrée de chaque cellule du second ordre. Pour tenir compte de la façon de réaliser cette limitation, on suppose que la puissance du bruit engendré est égale à k fois ($k \geqslant 1$) la puissance de bruit dû à une quantification à b_i bits. En effet cette limitation peut être un arrondi unique,

comme sur la figure 7.17 ou un arrondi après chaque multiplication comme assez souvent en pratique.

La détermination de la capacité des mémoires internes, et donc du nombre de bits des données b_i, doit tenir compte de la plage de variation des amplitudes. Si, dans la cascade de cellules du second ordre, on place en tête la cellule ayant le plus fort gain à la résonance H_m^1, alors le signal doit être divisé par H_m^1 à l'entrée du filtre pour garantir l'absence d'écrêtage, tout au moins pour les signaux sinusoïdaux, ce facteur H_m^1 étant donné par la relation (7.68). Il en résulte que la première cellule introduit une dégradation ΔSB du rapport signal à bruit qui s'exprime en décibels par :

$$\Delta SB = 10 \log \left[1 + (H_m^1)^2 \frac{k}{k_0} 2^{-2(b_i - b_d)} \right] \qquad (7.80)$$

La dégradation introduite par les autres cellules, qui ont une amplitude maximale plus faible est en général moins importante et peut être négligée. Il s'en suit que la capacité des mémoires peut être déterminée à partir de (7.80). Si la dégradation ΔSB doit être faible, alors :

$$b_i \simeq b_d + \frac{1}{2} \log 2 \left[4{,}3 . \frac{k}{k_0} . (H_m^1)^2 . \frac{1}{\Delta SB} \right] \qquad (7.81)$$

Compte tenu de la relation (7.568), en supposant $k = k_0$, on obtient :

$$b_i \simeq b_d + \frac{1}{2} \log 2 \left(\frac{1}{\Delta SB} \right) + \log 2 \left[\frac{f_e}{\Delta f} . \frac{1}{\sin\left(2\pi \frac{f_1}{f_e}\right)} \right] \qquad (7.82)$$

Dans cette expression la dégradation du rapport signal à bruit ΔSB est exprimée en décibels.

Cette expression est à rapprocher de la relation (5.53) pour les filtres RIF. Les filtres très sélectifs et à bande étroite demandent des nombres de bits élevés pour les mémoires internes.

Il est important de bien souligner que les évaluations ci-dessus, en particulier les relations (7.80) et (7.81) reposent sur une approche simplifiée. Dans une étude approfondie il faut calculer la dégradation ΔSB du rapport signal à bruit à la traversée du filtre en localisant exactement et en prenant en compte toutes les sources de bruit, et en adaptant en conséquence la relation (7.74) pour la détermination du bruit de calcul total. De plus il faut faire intervenir la puissance du signal en sortie du filtre puisque c'est en ce point que le bruit est calculé. La conception du filtre faisant intervenir une optimisation globale portant sur l'appairage des pôles et des zéros, l'ordre des cellules et les facteurs d'échelle représente en général un travail considérable qui ne se justifie que rarement.

Les résultats de ce paragraphe et du précédent sont illustrés par les deux exemples suivants :

Exemple 1

Soit le filtre du 4e ordre donné au paragraphe 7.2.4 (fig. 7.6) :

$$H_m^1 = 3{,}8 ; \quad H_m^2 = 2{,}2$$

on choisit :

$$a_0^0 = 0{,}25\,; \quad a_0^1 = 0{,}5 \quad \text{et} \quad a_0^2 = 0{,}549$$

La fonction de transfert finalement réalisée est la suivante :

$$\mathrm{H}(\mathrm{Z}) = 0{,}25\,.\,\frac{1 + 0{,}5974\,\mathrm{Z}^{-1} + \mathrm{Z}^{-2}}{1 - 0{,}670\,\mathrm{Z}^{-1} + 0{,}7144\,\mathrm{Z}^{-2}}\,.0{,}5\,\frac{1 + 1{,}632\,\mathrm{Z}^{-1} + \mathrm{Z}^{-2}}{1 - 0{,}814\,\mathrm{Z}^{-1} + 0{,}2636\,\mathrm{Z}^{-2}}\,.0{,}549$$

Le schéma du filtre est celui de la figure 7.21.

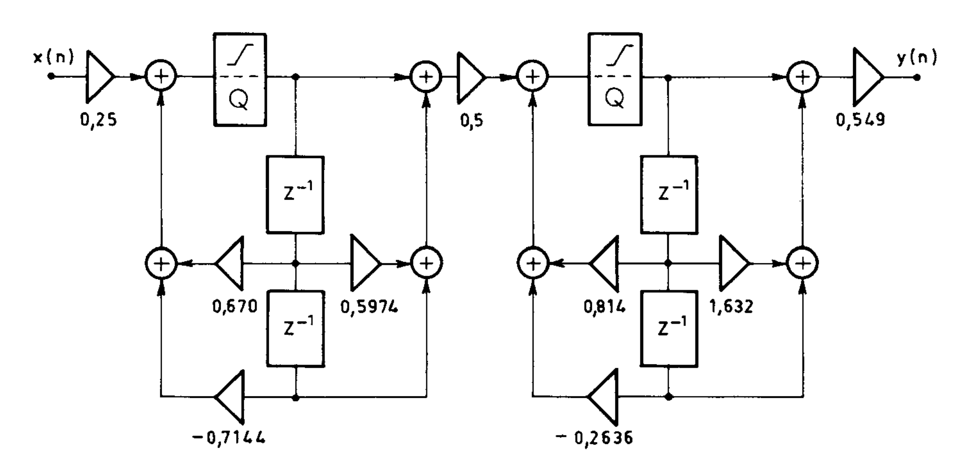


FIG. 7.21. *Schéma d'un filtre elliptique d'ordre 4*

Si les nombres d'entrée sont fournis par un convertisseur Analogique-Numérique à 12 bits et si les mémoires ont une capacité de 16 bits, la dégradation du rapport signal à bruit avec $k = 2$ et $k_0 = 1$ est estimée par (7.80) à :

$$\Delta SB = 10 \log\left(1 + \frac{1}{8}\right) \simeq 0{,}5 \text{ dB}$$

Exemple 2

Soit le filtre passe-bas très sélectif d'ordre 10 dont les coefficients ont pour valeur :

$a_1^1 = -1{,}41956$	$b_1^1 = -1{,}50269$	$b_2^1 = 0{,}98242$
$a_1^2 = -1{,}37231$	$b_1^2 = -1{,}49805$	$b_2^2 = 0{,}93652$
$a_1^3 = -1{,}22241$	$b_1^3 = -1{,}50916$	$b_2^3 = 0{,}85767$
$a_1^4 = -0{,}73120$	$b_1^4 = -1{,}53308$	$b_2^4 = 0{,}73730$
$a_1^5 = 1{,}07660$	$b_1^5 = -1{,}55640$	$b_2^5 = 0{,}62646$

et dont le gabarit correspond aux paramètres (fig. 5.7) :

$$\delta_1 = 0{,}01\,; \quad \delta_2 = 0{,}0002\,; \quad f_1 = 0{,}112\,; \quad f_2 = 0{,}117$$

Il vient :

$$H_m^1 = 87\,; \quad H_m^2 = 23\,; \quad H_m^3 = 12\,; \quad H_m^4 = 9\,; \quad H_m^5 = 15.$$

L'arrondi après multiplication par a_0^0, b_1^1 et b_2^1 conduit à prendre $k = 3$ dans l'estimation (7.80), qui, pour $k_0 = 1$ et $b_i = b_d$ conduit à : ΔSB = 43,5 dB.

Cette estimation peut être comparée à la solution optimale obtenue par programmation dynamique [13]. Dans cette solution, en désignant par Z_i le zéro correspondant au coefficient a_1^i et P_i le pôle correspondant à b_1^i et b_2^i $(1 \leqslant i \leqslant 5)$, les appairages et l'ordre des cellules sont les suivants :

$$P_4 - Z_2, \quad P_2 - Z_4, \quad P_5 - Z_1, \quad P_1 - Z_3, \quad P_3 - Z_5$$

et pour les facteurs d'échelle :

$$a_0^0 = 0{,}11838\,; \quad a_0^1 = 0{,}45112\,; \quad a_0^2 = 0{,}14724\,;$$
$$a_0^3 = 0{,}70252\,; \quad ,_0^4 = 0{,}43834\,; \quad a_0^5 = 0{,}38293$$

L'application de la relation (7.74) en prenant en compte toutes les sources de bruit, donne :

$$B_c = \frac{q^2}{12}\,.731$$

Si le signal d'entrée a un spectre uniforme, sa puissance en sortie est réduite par le facteur f_1/f_e. Il en résulte que la dégradation du rapport signal à bruit à la traversée du filtre s'écrit :

$$(\Delta SB)_{opt} = 34{,}7 \text{ dB}$$

Par rapport à l'approche simplifiée, dans le cas de ce filtre très sélectif, l'optimisation apporte un gain de 8,8 dB en bruit de calcul, soit, exprimé en nombre de bits des mémoires internes, un gain inférieur à 2 bits.

7.10 AUTO-OSCILLATIONS

En l'absence de signal à l'entrée du filtre RII, la limitation du nombre de bits des mémoires de données peut entraîner l'apparition d'auto-oscillations de faible amplitude et de forte amplitude.

Des auto-oscillations peuvent se produire aux grandes amplitudes, par dépassement de la capacité des mémoires. L'équation du système s'écrit alors :

$$y(n) + \sum_{i=1}^{N} b_i\, y(n-i) = 0 \tag{7.83}$$

La condition d'absence naturelle de tels phénomènes s'exprime par l'inégalité :

$$\left| \sum_{i=1}^{N} b_i y(n-i) \right| < 1$$

Pour des filtres d'ordre supérieur à 2, on montre que la présence d'un dispositif de saturation logique ne suffit plus à garantir l'absence d'auto-oscillations de grande amplitude [14]. Par contre un filtre réalisé par une cascade de cellules du second ordre avec dispositif de saturation logique ne présente pas cette possibilité.

Les auto-oscillations de faible amplitude produites par une cellule, se trouvent filtrées par les cellules suivantes, pour lesquelles le signal d'entrée n'est pas nul.

Si la stratégie qui consiste à interconnecter les cellules par ordre de maximum décroissant est appliquée, en l'absence de signal à l'entrée du filtre, la première cellule peut produire une auto-oscillation, à une fréquence voisine de la fréquence de résonance c'est-à-dire à la limite de la bande passante pour un filtre très sélectif. En fait l'auto-oscillation correspond à l'insertion dans la chaîne de la figure 7.21 d'un signal parasite e_0 dont l'amplitude est limitée par la quantification comme indiqué au paragraphe VI.7. L'amplitude A_a de l'auto-oscillation à la sortie du filtre de gain unité et en structure cascade D-N peut ainsi être estimée par

$$A_a \simeq H_m^1 . 2^{-b_i}$$

où b_i désigne comme précédemment le nombre de bits des mémoires internes.

D'après la relation (7.68) il vient :

$$A_a \simeq 2^{-b_i-1} \frac{f_e}{\Delta f} \cdot \frac{1}{\sin\left(2\pi \frac{f_1}{f_e}\right)} \qquad (7.84)$$

Cette expression fournit une relation entre l'amplitude des auto-oscillations et les caractéristiques du filtre, pour la méthode de réalisation considérée.

D'autres méthodes de réalisation, par exemple l'interconnexion des cellules dans un ordre différent, peuvent conduire à des valeurs plus faibles de l'amplitude de ces signaux parasites.

7.11 COMPARAISON ENTRE LES FILTRES RII ET RIF

Les deux types de filtres examinés RII et RIF, permettent de satisfaire un gabarit quelconque donné. La question du choix entre ces deux approches se pose fréquemment au concepteur de systèmes. Le critère est la complexité des circuits à mettre en œuvre ; en pratique la comparaison se ramène principalement à l'évaluation du paramètre simple que constitue le nombre de multiplications à effectuer comme on le verra dans un chapitre ultérieur.

Les relations (5.32) et (7.29) donnent des estimations de l'ordre N des filtres RIF et RII nécessaire pour satisfaire des spécifications de filtrage passe-bas exprimées par l'ondulation en bande passante et affaiblie et par la largeur de bande passante et de bande de transition.

Pour le filtre RIF à phase linéaire les coefficients présentent une symétrie et, si N est pair, $\frac{N}{2}$ ont des valeurs différentes. Un tel filtre nécessite donc $\frac{N}{2}$ mémoires de coefficients et N mémoires de données internes. Pour chaque nombre de sortie il faut faire $\frac{N}{2}$ multiplications et N additions. Si la linéarité en phase n'est pas imposée, il est possible de réduire l'ordre N, avec les filtres à déphasage minimal, comme l'indique la relation (5.62). Cette réduction dépend de l'ondulation en bande passante et reste inférieure à 50 % ; il s'en suit une augmentation du nombre de multiplications puisque la symétrie des coefficients disparaît. Parmi les avantages des filtres RIF, il faut souligner qu'ils sont toujours stables et qu'ils sont faciles à réaliser.

Les filtres RII sont plus délicats à mettre en œuvre ; sauf cas particulier, c'est le type elliptique qui est le plus efficace et le plus utilisé. L'estimation (7.29) montre que l'ordre du filtre passe par un maximum au voisinage de la fréquence $\frac{f_e}{4}$, comme le montre la figure 7.22. Soit n cet ordre ; en supposant que le filtre est réalisé en cellules du second ordre comportant chacune 4 coefficients et 2 mémoires de données (paragraphe 6.4), il faut pour réaliser le filtre n mémoires de données, $2n$ mémoires de coefficients et chaque nombre de sortie exige $2n$ additions et $2n$ multiplications.

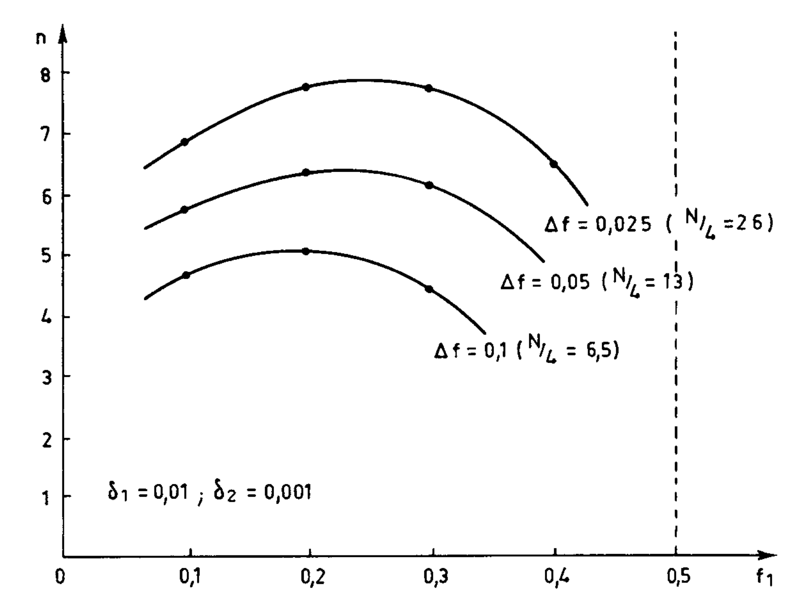


FIG. 7.22. *Ordre d'un filtre passe-bas elliptique en fonction de la largeur de la bande passante*

Si la comparaison entre filtres RIF et RII se limite au nombre de multiplications à faire pour obtenir un nombre en sortie, dans le cas du passe-bas, le type RII est plus avantageux que le type RIF pour les valeurs de paramètres telles que :

$$N > 4n \tag{7.85}$$

D'après les relations (5.32) et (7.29) c'est la bande de transition qui est le paramètre le plus important dans la comparaison et l'on a à peu près le même nombre de multiplications, dans les conditions les plus défavorables pour le filtre RII, pour une bande de transition telle que :

$$\frac{1}{3}\,\frac{f_e}{\Delta f} \simeq 2\log\left(\frac{f_e}{\Delta f}\right)$$

c'est-à-dire $\Delta f \simeq f_e/3$. Il s'en suit que l'inégalité (7.85) est vérifiée dès que Δf est inférieur à $f_e/3$. C'est le cas dans la grande majorité des applications. Par exemple, pour les valeurs de paramètres correspondant à la figure 7.22, cette inégalité est toujours vérifiée.

La linéarité en phase peut être approchée dans une bande de fréquence limitée avec un filtre RII, en complétant par exemple le filtre elliptique de base par un ensemble de circuits d'égalisation du temps de propagation de groupe. Ces circuits sont des déphaseurs purs dont les propriétés ont été exposées au paragraphe VI.3. L'expérience montre que le filtre RIF, qui présente une linéarité en phase parfaite, demande toujours moins de calculs [15] ; il est, de plus, facile à réaliser.

Finalement, il est recommandé d'utiliser les filtres RIF quand la linéarité en phase est demandée et les filtres RII dans les autres cas.

Cependant, la comparaison ci-dessus a été faite avec l'hypothèse implicite que la cadence à laquelle se présentent les nombres, est la même en entrée et sortie du filtre. Les termes de la comparaison se trouvent sensiblement modifiés si cette contrainte disparaît, comme le montre un chapitre ultérieur.

BIBLIOGRAPHIE

[1] A. Oppenheim and R. Schafer – *Digital Signal Processing*. Chapters 5 and 9, Prentice Hall, New Jersey 1974.

[2] L. Rabiner and B. Gold – *Theory and Application of Digital Signal Processing.* Chapters 4 and 5, Prentice Hall, New Jersey 1975.

[3] R. Boite et H. Leich – *Les filtres Numériques. Analyse et synthèse des filtres unidimensionnels.* Collection CNET-ENST, Éditions Masson 1980.

[4] B. Gold and C. Rader – *Digital Processing of Signals*. Mac Graw Hill New York 1969.

[5] A. G. Constantinides – Spectral Transformation for Digital Filters. *Proc. IEE,* N° 8, 1970.

[6] A. Deczky – Synthesis of Recursive Digital Filters Using the Minimum *p*-error Criterion. *IEEE Trans. On Audio and Electroacoustics,* Vol. AU20, Oct. 1972.

[7] R. Fletcher and M. J. D. Powell – A Rapidly Convergent Descent Method for Minimization. *Computer J.,* 6, N° 2, 1963.

[8] J. P. Thiran – Equal Ripple Delay Recursive Filters. *IEEE Trans. On Circuit Theory,* November 1971.

[9] T. DURRANI and R. CHAPMAN – « Optimal All-pole Filter Design Based on Discrete Prolate Spheroidal sequences », *IEEE Trans.,* Vol. ASSP-32, N° 4, August 1984, pp. 716-721.

[10] P. P. VAIDYANATHAN, S. K. MITRA and Y. NEUVO – « A New Approach to the Realisation of Low Sensitivity IIR Filters », *IEEE Trans.,* Vol. ASSP-34, N° 2, April 1986, pp. 350-361.

[11] L. B. JACKSON – Round-off Noise Analysis for Fixed Point Digital Filters in Cascade or Parallel form. *IEEE on Audio and Electroacoustics,* June 1970.

[12] A. PELED and B. LIU – *Digital Signal Processing : Theory, Design and Implementation.* New York, John Wiley, 1976.

[13] Von E. LUEDER, H. HUG and W. WOLF – *« Minimizing the Round-off Noise in Digital Filters by Dynamic Programming »,* Frequenz, Vol. 29, N° 7, 1975, pp. 211-214.

[14] D. MITRA – Large Amplitude, Self Sustained oscillations in Difference Equations, which describe Digital Filter Sections Using Saturation Arithmetic. *IEEE Trans.* Vol. ASSP-25, N° 2, April 1977.

[15] L. RABINER, J. F. KAISER, O. HERRMANN and M. DOLAN – Some Comparison between FIR and IIR Digital Filters. *BSTJ,* Vol. 53, Feb. 1974.

EXERCICES

1 Utiliser les formules du paragraphe 7.1 pour calculer la réponse en fréquence et en phase et le temps de propagation de groupe de la cellule définie par la relation :

$$y(n) = x(n) + 0{,}7\,x(n-1) + 0{,}9\,y(n-1)$$

Même question pour la cellule du second ordre de fonction de transfert en Z :

$$H(Z) = \frac{b_2 + b_1 Z^{-1} + Z^{-2}}{1 + b_1 Z^{-1} + b_2 Z^{-2}}$$

2 Pour calculer un filtre numérique passe-bande on se propose d'utiliser des abaques pour filtres analogiques. De quel gabarit doit-on partir pour que le filtre numérique affaiblisse dans les bandes (0 – 0,15) et (0,37 – 0,5) et ne présente pas d'affaiblissement dans la bande (0,2 – 0,33) en supposant $f_e = 1$? Étudier le calcul direct à partir de la transformation d'un passe-bas par la relation 7.31.

3 Calculer les coefficients d'un filtre de Butterworth d'ordre 4 dont l'amplitude prend la valeur $\frac{1}{\sqrt{2}}$ à la fréquence $f_c = 0{,}25$. Donner la décomposition en cellules du second ordre.

4 Utiliser une transformation en fréquence pour transformer le filtre passe-bas du paragraphe 7.2.3 (fig. 7.6) en passe haut avec pour limite de bande passante $f_H = 0{,}4$. Comment évoluent les pôles et les zéros dans cette opération ?

5 Donner la décomposition en cellules du second ordre du filtre de la figure 7.10. Calculer les facteurs d'échelle des cellules et la longueur des mémoires si le bruit de calcul ajouté par le filtre doit rester inférieur en puissance à $\frac{1}{10}$ du bruit présent à l'entrée et si les nombres à l'entrée comptent 10 bits.

6 Le gabarit de la figure 7.10 est élargi de 0,1 dB pour permettre un arrondi des coefficients du filtre. Combien de bits sont nécessaires pour représenter les coefficients dans la structure cascade ? Faire l'évaluation pour la structure parallèle. Rechercher un optimum pour l'arrondi des coefficients ; peut-on réduire le nombre de bits trouvé précédemment ?

7 Le filtre donné en exemple au paragraphe 7.2.2 présente-t-il des auto-oscillations ? Quelles sont les fréquences et les amplitudes ? Même question pour le filtre de la figure 7.20.

8 Quelle est la quantité de calculs demandée par le filtre de la figure 7.6 ? Combien de mémoires sont nécessaires ? Combien de coefficients demande un filtre RIF pour le même gabarit. Comparer les quantités de calculs et les capacités de mémoire.

9 Dans un équipement de transmission numérique MIC on se propose de réaliser la fonction de filtrage de voie par technique numérique.

Le signal téléphonique est échantillonné à 32 KHz et codé à 12 bits, le filtrage est effectué par un filtre RII de type passe-bas.

La bande passante est 3 300 Hz, la bande affaiblie commence à 4 600 Hz.

Les ondulations en bandes passante et affaiblie ont les valeurs :

$$\delta_1 \leqslant 0{,}015 ; \quad \delta_2 \leqslant 0{,}04$$

Le programme de calcul des filtres elliptiques fournit les résultats suivants :

ordre du filtre : N = 4.

Zéros : $Z_1 = 0{,}09896 \pm j\,0{,}995$

$Z_2 = 0{,}5827 \pm j\,0{,}8127$

Pôles : $P_1 = 0{,}6192 \pm j\,0{,}2672$

$P_2 = 0{,}702 \pm j\,0{,}589$

- Calculer la fonction de transfert du filtre décomposé en cellules du 2nd ordre.
- Quelle est la valeur du facteur d'échelle global sachant que l'amplitude à la fréquence 0 est 0,99.
- Les coefficients sont quantifiés à 10 bits. Déterminer le déplacement des pointes infinies et évaluer le supplément d'ondulation en bande passante.
- Calculer le facteur d'échelle à affecter à chaque cellule et estimer le bruit de calcul produit si les mémoires de données ont 16 bits.
- Donner le schéma complet du filtre.
- Évaluer la complexité en :

Nombre de multiplications et additions par seconde.

Nombre de bits de mémoires.

10 On considère le gabarit de filtre suivant : BP = 0,2 dB ; BA = 45 dB ; FB = 1700 H_2 ; FA = 2 000 H_2 ; FE = 8 000 H_2.

Calculer l'ordre du filtre nécessaire. En prenant comme ordre N = 6, on obtient une marge sur les ondulations ; calculer cette marge et comparer le résultat à ce qui est obtenu avec les pôles et zéros fournis par un programme de calcul, en traçant la réponse en fréquence.

Pôles	Zéros
0,195886 ± j 0,926044	– 0,210790 ± j 0,999778
0,269059 ± j 0,736096	– 0,235254 ± j 0,971934
0,394027 ± j 0,307068	– 0,814637 ± j 0,579971

Calculer le nombre de bits des coefficients.

La dégradation du rapport signal à bruit à la traversée du filtre étant limitée à 0,1 dB, quelle est l'augmentation du nombre de bits des mémoires internes par rapport au nombre de bits des nombres représentant le signal d'entrée.

Donner le schéma complet du filtre.

Chapitre 8

Les structures de filtres en chaîne

Les structures de filtres présentées dans les précédents chapitres se déduisent directement de la fonction de transfert en Z de ces filtres; les coefficients appliqués aux circuits multiplieurs sont les coefficients des puissances de Z^{-1}. Avec des opérations supplémentaires à partir de la fonction de transfert, on peut aboutir à des structures plus élaborées ayant des propriétés intéressantes, c'est le cas des structures en chaîne.

En filtrage analogique, les structures en chaîne permettent de réaliser des filtres ayant des ondulations très faibles et une excellente sélectivité, avec des composants passifs de précision limitée. En numérique ces propriétés peuvent se traduire par une réduction du nombre de bits à affecter à la représentation des coefficients et, par suite, des gains en circuits et une réduction du bruit de calcul.

Les filtres analogiques en chaîne sont basés sur la mise en cascade de quadripôles dont les propriétés vont d'abord être rappelées [1].

8.1 PROPRIÉTÉS DES QUADRIPÔLES

Le quadripôle général fermé sur les résistances R1 et R2 est présenté sur la figure 8.1 avec les variables courant I et tension V aux accès 1 et 2. Ce quadripôle supposé linéaire est défini par sa matrice d'impédance z, qui traduit les relations entre les variables, généralement prises sous forme réduite avec :

$$R = R1 = R2; \qquad v = \frac{V}{\sqrt{R}}; \quad i = I.\sqrt{R}$$

FIG. 8.1. *Quadripôle avec terminaisons résistives*

Il vient :

$$v = zi \tag{8.1}$$

avec :

$$z = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix}$$

Les valeurs z_{12} et z_{21} sont les impédances de transfert du quadripôle. Il est dit réciproque si : $z_{12} = z_{21}$.

Si en retournant le quadripôle, le régime extérieur n'est pas modifié, il est dit symétrique et l'on a : $z_{11} = z_{22}$.

Pour faire apparaître les coefficients de transmission et réflexion du quadripôle, on le définit par une autre matrice, la matrice de répartition. En se plaçant dans une situation de référence où les résistances de terminaison sont unitaires, on considère les ondes incidentes et réfléchies a et b telles que :

$$a = \frac{1}{2}(v + i) \tag{8.2}$$

$$b = \frac{1}{2}(v - i) \tag{8.3}$$

Les variables a et b sont liées par des relations qui s'obtiennent en utilisant (8.1) :

$$a = \frac{1}{2}(z + \mathrm{I}_2)i \quad ; \quad \mathrm{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$b = \frac{1}{2}(z - \mathrm{I}_2)i$$

Il vient :

$$b = \mathrm{S}a \tag{8.4}$$

avec :

$$\mathrm{S} = \begin{bmatrix} \mathrm{S}_{11} & \mathrm{S}_{12} \\ \mathrm{S}_{21} & \mathrm{S}_{22} \end{bmatrix}$$

et :

$$\mathrm{S} = (z - \mathrm{I}_2)(z + \mathrm{I}_2)^{-1} \tag{8.5}$$

Si le quadripôle est réciproque, alors on a :

$$\mathrm{S}_{12} = \mathrm{S}_{21} = \tau \tag{8.6}$$

où τ est le coefficient de transmission tel que :

$$\tau = \frac{2\mathrm{V}_2}{\mathrm{E}} \tag{8.7}$$

En désignant par z_1 et z_2 les impédances vues à l'entrée et à la sortie du quadripôle respectivement, il vient :

$$S_{11} = \rho_1 = \frac{z_1 - 1}{z_1 + 1}\ ; \quad S_{22} = \rho_2 = \frac{z_2 - 1}{z_2 + 1} \tag{8.8}$$

Les valeurs ρ_1 et ρ_2 sont les coefficients de réflexion à l'entrée et à la sortie du quadripôle.

Si le quadripôle est non dissipatif, la puissance active qu'il absorbe est nulle. On montre que la matrice de répartition d'un quadripôle réciproque non dissipatif prend la forme canonique suivante [2] :

$$S = \frac{1}{g}\begin{bmatrix} h & f \\ f & \pm h_* \end{bmatrix} \tag{8.9}$$

où f, g et h sont des polynômes réels ayant les propriétés suivantes :

– Ils sont liés par une relation, qui sur l'axe imaginaire correspond à :

$$|g|^2 = |h|^2 + |f|^2$$

La notation $h_*(p)$ correspond à $h(-p)$.

– Suivant que f est de degré pair ou impair le signe inférieur ou supérieur est à prendre dans (8.9).

– Toutes les racines de g dans le plan complexe sont dans le demi-plan de gauche.

Les polynômes f, g, h sont les polynômes caractéristiques du quadripôle ; les racines de $f(p)$ sont en général sur l'axe imaginaire en bande affaiblie, ce sont les zéros de transmission. Les racines de $h(p)$ sont les zéros d'affaiblissement et pour un quadripôle non dissipatif ils se trouvent en général sur l'axe imaginaire dans la bande passante.

Pour le quadripôle de la figure 8.1, le coefficient de transmission s'écrit :

$$S_{12} = \frac{2V_2}{E}\sqrt{\frac{R_2}{R_1}} \tag{8.10}$$

On désigne par affaiblissement la fonction $A_f(\omega)$ exprimée en dB et définie par :

$$A_f(\omega) = -10 \log |S_{12}(\omega)|^2 = 10 \log \left|\frac{g(\omega)}{f(\omega)}\right|^2 = 10 \log \left(1 + \frac{|h|^2}{|f|^2}\right)$$

La relation suivante :

$$\left|\frac{f(\omega)}{g(\omega)}\right|^2 + \left|\frac{h(\omega)}{g(\omega)}\right|^2 = 1 \tag{8.11}$$

exprime simplement le fait que la puissance non transmise est réfléchie.

Pour la mise en cascade de quadripôles il est intéressant de considérer également la matrice de transfert t définie par :

$$\begin{bmatrix} b_1 \\ a_1 \end{bmatrix} = t \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \tag{8.12}$$

La mise en cascade de quadripôles se traduit par le produit de leurs matrices de transfert.

La matrice de transfert des quadripôles non dissipatifs se met sous la forme canonique suivante :

$$t = \frac{1}{f} \begin{bmatrix} \pm g_* & h \\ \pm h_* & g \end{bmatrix} \tag{8.13}$$

FIG. 8.2. *Matrices de transfert de quadripôles élémentaires*

A titre d'exemple la figure 8.2 donne les matrices de transfert de quelques quadripôles élémentaires.

Le fait que le quadripôle soit non dissipatif a des conséquences importantes sur l'affaiblissement $A_f(\omega)$. En effet, en bande passante, la fonction $A_f(\omega)$ ne peut prendre de valeurs négatives et par suite aux fréquences où $h(\omega)$ s'annule, il doit en être de même de la dérivée de $A_f(\omega)$ par rapport à l'un quelconque des paramètres. Dans un filtre à inductances et capacités terminé par des résistances, une variation des valeurs des éléments L et C n'affecte pas l'affaiblissement au premier ordre, aux fréquences où il s'annule.

Si l'ondulation est faible on peut supposer que cette propriété s'étend à toute la bande passante. Pratiquement on peut considérer que, dans un filtre en échelle par exemple, les interactions entre les différentes branches sont telles qu'une dérive sur un élément se répercute sur l'ensemble des facteurs de la fonction d'affaiblissement avec un effet de compensation globale qui en minimise l'incidence.

Dans ces conditions, il apparaît très intéressant de rechercher des structures de filtres numériques ayant des propriétés similaires. En effet dans un filtre numérique dont les amplitudes d'ondulation en bandes passante et affaiblie sont comparables, c'est le dénominateur de la fonction de transfert qui fixe le nombre de bits nécessaire à la représentation des coefficients. On peut donc espérer un gain important avec les structures déduites par exemple des filtres en échelle, sur les coefficients eux-mêmes, sur la complexité des multiplieurs et également sur la puissance de bruit de calcul.

Les structures en échelle sont les plus couramment utilisées en filtrage analogique passif. La procédure pour obtenir les éléments d'une telle structure à partir d'une fonction de transfert donnée est décrite en détail dans la référence [2]. Elle consiste à factoriser la matrice de transfert globale, définie à partir de la fonction de transfert $H(\omega)$ calculée, en matrices partielles correspondant aux bras série et parallèle de la structure en échelle.

L'approche la plus directe pour obtenir une structure de filtre numérique à partir d'une structure de filtre analogique en échelle consiste à simuler le graphe de fluence dans le domaine tension-courant.

8.2 LES FILTRES EN ÉCHELLE SIMULÉE

La représentation des filtres en échelle à l'aide de leur graphe de fluence est utilisée pour la synthèse des filtres actifs à l'aide de circuits intégrateurs ou différenciateurs.

Pour mettre en évidence le graphe de fluence dans le domaine tension-courant, on considère le filtre en échelle de la figure 8.3, terminé sur les résistances R1 et R2.

FIG. 8.3. *Filtre en échelle analogique*

L'application des lois de Kirchhoff conduit aux relations :

$$I_1 = (E - V_2)R_1^{-1}; \quad I_{K-1} = (V_{K-2} - V_K)Z_{K-1}^{-1}; \quad I_{N+1} = V_N R_2^{-1}$$
$$V_2 = (I_1 - I_3)Y_2^{-1}; \quad V_K = (I_{K-1} - I_{K+1})Y_K^{-1}$$

où l'indice K prend les valeurs : 4, 6, …, N.

Le graphe de fluence est composé d'arcs à chacun desquels est associé un coefficient représentant une impédance ou une admittance. A chaque sommet est associé soit la tension en un nœud soit le courant dans une branche. Pour chaque arc on forme le produit du coefficient correspondant par la grandeur associée à son origine et la grandeur attachée à chaque sommet est la somme des produits correspondant aux différents arcs incidents.

Le graphe du filtre en échelle de la figure 8.3 est représenté sur la figure 8.4. Les sommets auxquels sont associés les courants et les tensions alternativement, se succèdent. La topologie ainsi obtenue est dite à « boucles imbriquées », et dans la littérature anglaise « leapfrog ».

FIG. 8.4. *Graphe de fluence d'un filtre en échelle*

Les filtres numériques dits en échelle simulée sont des structures obtenues en simulant chaque arc du groupe ou chaque branche de l'échelle par un organe de fonction de transfert équivalente.

Un cas particulièrement simple est celui où les impédances séries Z_{K-1} sont des inductances et les branches parallèles Y_K sont des capacités (K = 4, 6, …, N). Ce cas est celui des filtres sans pointes d'affaiblissement infini ou purement récursifs. Les fonctions de transfert à prendre en compte sont de la forme :

$$Z_{K-1}^{-1} = \frac{R}{sL_{K-1}} \; ; \quad Y_K^{-1} = \frac{1}{sC_K R} \tag{8.14}$$

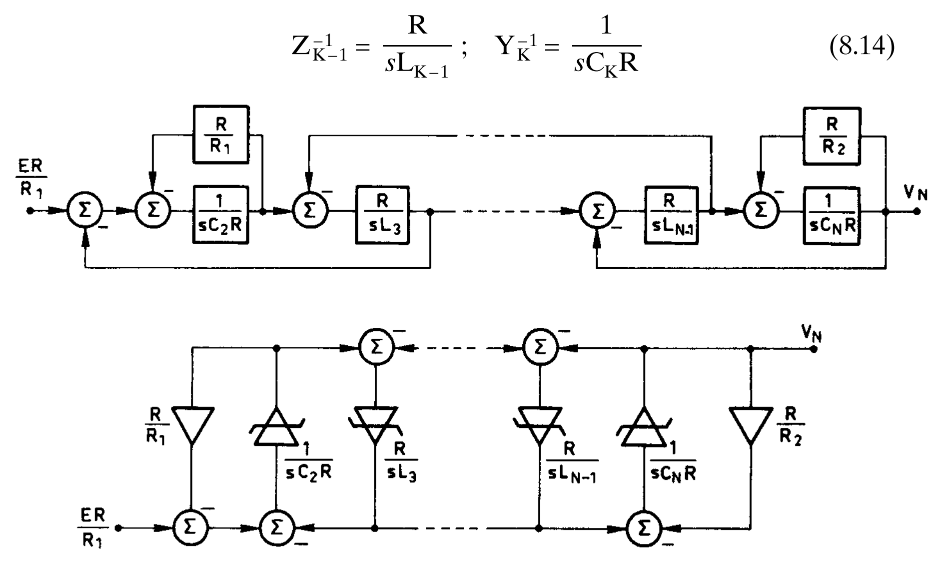


FIG. 8.5. *Schéma de filtre réalisé par intégrateurs*

où s est la variable de Laplace et R une constante de normalisation. Il s'agit dans les deux cas de fonctions de transfert d'intégrateurs aisément réalisables avec des amplificateurs opérationnels et des réseaux R – C. On aboutit alors au diagramme de la figure 8.5, déduit de la figure 8.4, qui représente les fonctions réalisées et le schéma du circuit avec les intégrateurs.

La réalisation numérique consiste à remplacer chaque intégrateur par une fonction équivalente. Dans la référence [3] il apparaît que le seul circuit numérique intégrateur simple réalisable et équivalent à un intégrateur analogique est celui qui est représenté par la figure 8.6 et dont la fonction de transfert en Z, I(Z) s'écrit :

$$I(Z) = \frac{aZ^{-\frac{1}{2}}}{1 - Z^{-1}} \tag{8.15}$$

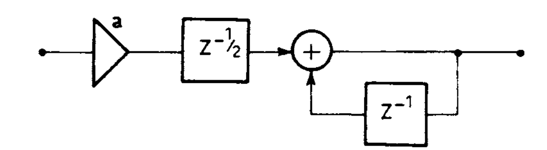


FIG. 8.6. *Circuit numérique intégrateur*

L'équivalence entre les intégrateurs analogiques et numériques est obtenue comme pour toute fonction de transfert, en remplaçant Z par $e^{j\omega T}$. Il vient :

$$I(\omega) = \frac{a e^{-j\omega\frac{T}{2}}}{1 - e^{-j\omega T}} = \frac{a}{2j} \cdot \frac{1}{\sin\left(\omega \frac{T}{2}\right)} \tag{8.16}$$

T est la période d'échantillonnage du circuit numérique. Cette fonction est équivalente à la fonction $\frac{a}{j\omega T}$ analogique avec une transformation de l'échelle des fréquences. Si f_A désigne la fréquence analogique et f_N la fréquence numérique on a :

$$\pi f_A T = \sin(\pi f_N T) \tag{8.17}$$

La déformation ainsi apportée à l'échelle des fréquences est différente de celle obtenue avec la transformation bilinéaire introduite au chapitre précédent, comme le montre la figure 8.7. Il faut tenir compte de cette déformation dans le calcul d'un filtre à partir d'un gabarit.

Le circuit de la figure 8.6 présente l'inconvénient de faire apparaître la fonction $Z^{-\frac{1}{2}}$ qui correspond à un circuit de mémoire supplémentaire. Or la fonction de transfert d'un filtre en échelle n'est pas modifiée quand les impédances de toutes les branches sont multipliées par une même fonction [3]. Cette propriété

Fig. 8.7. *Déformation de l'échelle des fréquences par transformation sinusoïdale*

a déjà été utilisée précédemment pour introduire la constante de normalisation R.

En multipliant les impédances par $Z^{-\frac{1}{2}}$, on élimine ce terme de toutes les fonctions de transfert en Z des intégrateurs du circuit qui deviennent :

$$I_i(Z) = \frac{TR}{L_i} \cdot \frac{1}{1 - Z^{-1}} \qquad \text{pour } i \text{ impair}$$

et :

$$I_i(Z) = \frac{T}{C_i R} \cdot \frac{Z^{-1}}{1 - Z^{-1}} \qquad \text{pour } i \text{ pair}$$

Par contre les résistances de terminaison sont transformées en $R_1 Z^{-\frac{1}{2}}$ et $R_2 Z^{-\frac{1}{2}}$; les terminaisons ne sont plus purement résistives, elles ont les fonctions de transfert :

$$R_1 e^{-j\pi f T}; \qquad R_2 e^{-j\pi f T}$$

Quand la fréquence d'échantillonnage est grande devant la bande passante cet effet peut être négligé ; il en résulte une modification non significative de la fonction de transfert du filtre. De plus les résistances R_1 et R_2 peuvent être choisies unitaires, de même que la constante de normalisation R. Le schéma du filtre numérique obtenu dans ces conditions est donné par la figure 8.8.

Les coefficients ont les valeurs suivantes, pour un ordre N impair :

$$a_N = \frac{T}{C_{N+1}}; \quad a_{2i-1} = \frac{T}{C_{2i}}; \quad a_{2i} = \frac{T}{L_{2i+1}}; \quad i = 1, 2, \dots \frac{N-1}{2} \tag{8.18}$$

Le filtre ainsi réalisé nécessite N multiplications et N mémoires pour une fonction de transfert d'ordre N ; pour ces paramètres la structure est canonique. Le nombre d'additions s'élève à 2N + 1.

FIG. 8.8. *Filtre numérique en échelle simulée*

En résumé le calcul d'un filtre numérique en échelle simulée à partir d'un gabarit imposé demande les étapes suivantes :

– Transposer le gabarit en modifiant l'échelle des fréquences à l'aide de la relation (8.17) ci-dessus.

– Calculer les éléments d'un filtre à éléments passifs LC en échelle, satisfaisant au gabarit transposé.

– A partir des valeurs d'éléments obtenues calculer les coefficients a_i ($i = 1, 2, \text{N}$) du filtre numérique par les relations (8.18).

L'intérêt principal de la structure obtenue est que les coefficients peuvent être représentés par un nombre de bits très faible, qui peut ne pas dépasser quelques unités ; alors certaines multiplications peuvent se ramener à de simples additions et l'on peut même dans certains cas éliminer toutes les multiplications du filtre, ce qui se traduit par une économie substantielle de circuits. Pour illustrer cette propriété, on considère le filtre passe-bas d'ordre N = 7 [3], dont les éléments ont pour valeur (fig. 8.5) :

$$\begin{aligned}
&\text{R} = \text{R}_1 = \text{R}_9 = 1\\
&\text{C}_2 = 1{,}2597 = \text{C}_8\\
&\text{L}_3 = 1{,}5195 = \text{L}_7\\
&\text{C}_4 = 2{,}2382 = \text{C}_6\\
&\text{L}_5 = 1{,}6796
\end{aligned}$$

Les coefficients a_i ($i = 1, 2, \ldots, \text{N}$) du filtre numérique en échelle simulée correspondant sont calculés avec une période d'échantillonnage $\text{T} = \dfrac{1}{f_e} = 0{,}01$ par les formules (8.18) ci-dessus. Les ondulations du filtre en bande passante sont reportées sur la figure 8.9, quand les coefficients sont représentés par 10, 5 et 3 bits. Il est remarquable de constater qu'avec la représentation à 5 bits, les zéros d'affaiblissement sont conservés. Avec 3 bits, ils sont conservés également à l'exception de celui qui est le plus proche de la bande de transition. Ainsi se trouve vérifiée la propriété d'insensibilité au premier ordre des zéros d'affaiblissement, énoncée au paragraphe 8.1.

Fig. 8.9. *Ondulations en bande passante pour diverses représentations des coefficients*

Par rapport à la structure cascade du chapitre précédent, le gain obtenu sur cet exemple peut être estimé à 4 ou 5 bits pour la représentation des coefficients.

La technique décrite dans ce paragraphe peut s'étendre aux types de filtres autres que le passe-bas purement récursif, mais avec des complications de schémas. D'autre part la nécessité d'avoir une fréquence d'échantillonnage grande devant la bande passante n'est pas très favorable pour l'efficacité du traitement. En fait la structure en échelle simulée est surtout utilisée avec un autre mode de réalisation des calculs, celui qui apparaît dans les dispositifs à commutation de capacités.

8.3 LES DISPOSITIFS À COMMUTATION DE CAPACITÉS (DCC)

Les filtres utilisant les dispositifs à commutation de capacités ne sont pas des filtres numériques au sens strict, car ils ne font pas appel aux opérateurs arithmétiques. Néanmoins ils emploient les mêmes méthodes de calcul et sont complémentaires des filtres numériques. Ils sont très utilisés dans les conversions analogique-numérique.

Le principe de base, qui a été introduit dans la référence [4], est le suivant : le fait de commuter une capacité C entre deux tensions V_1 et V_2 à la fréquence f_e est équivalent à introduire une résistance R telle que :

$$R = \frac{1}{C} f_e$$

entre les deux tensions. En effet, comme le montre la figure 8.10, la capacité se charge sous les tensions V_1 et V_2 alternativement et il s'en suit un transfert de

charge $C(V_1 - V_2)$; si les opérations sont effectuées à la cadence f_e, le courant i tel que :

$$i = C(V_1 - V_2) f_e$$

circule entre les tensions V_1 et V_2.

FIG. 8.10. *Commutation d'une capacité entre les tensions V_1 et V_2*

Cette résistance équivalente s'introduit dans un circuit intégrateur comme indiqué sur la figure 8.11. L'intégrateur considéré présente un sommateur à son entrée, comme ceux de la figure 8.5. L'équation décrivant le fonctionnement de l'intégrateur analogique est donnée. Dans le schéma avec commutation de capacité, la capacité C_1 est appliquée, à la cadence f_e, alternativement entre les tensions V_0 et V_1 d'une part et l'entrée de l'amplificateur opérationnel d'autre part. L'expression de la variation ΔV_2 de la tension de sortie pendant la durée Δt, supposée grande devant la période $\frac{1}{f_e}$, est indiquée sur le schéma.

FIG. 8.11. *Intégrateur à commutation de capacités*

La condition d'équivalence entre les deux types d'intégrateurs s'écrit :

$$C_1 = \frac{1}{f_e R} \tag{8.19}$$

Cependant pour analyser complètement l'intégrateur à commutation de capacités il faut tenir compte de l'échantillonnage [5] et calculer sa fonction de transfert en Z. Soit $v_e(t)$ le signal d'entrée et $v_2(t)$ le signal de sortie ; la période d'échantillonnage T est supposée divisée en deux parties égales. La capacité C_1 est connec-

tée pendant $\frac{T}{2}$ à l'entrée de l'intégrateur et pendant $\frac{T}{2}$ la tension $v_e(t)$ lui est appliquée ; supposons que ce soit entre les instants nT et $\left(n+\frac{1}{2}\right)$T. La charge transmise à l'intégrateur s'écrit Q(nT) telle que :

$$Q(nT) = C_1 v_e\left[\left(n+\frac{1}{2}\right)T\right]$$

Dans ces conditions à l'instant $(n + 1)$T, la tension en sortie s'écrit :

$$v_2[(n+1)T] = V_2(nT) - \frac{C_1}{C_2}\, v_e\left[\left(n+\frac{1}{2}\right)T\right]$$

En prenant la transformée en Z des deux membres, il vient :

$$\frac{V_2(Z)}{V_e(Z)} = H(Z) = -\frac{C_1}{C_2}\cdot\frac{Z^{-\frac{1}{2}}}{1-Z^{-1}} \qquad (8.20)$$

On retrouve le type de fonction de transfert donnée par la relation (8.15) pour les circuits numériques ; l'intégrateur à commutation de capacités réalise exactement les mêmes fonctions que les circuits numériques du paragraphe précédent et la même déformation de l'axe des fréquences intervient. Il faut noter que, pour qu'aucun retard supplémentaire ne s'introduise et pour que cette fonction soit conservée dans la mise en cascade de deux intégrateurs, les capacités de ces deux intégrateurs doivent être commutées en opposition de phase.

Le schéma avec dispositif à commutation de capacités d'un filtre en échelle simulée, comme celui de la figure 8.5 est obtenu par substitution de circuits intégrateurs, en calculant dans chaque cas la valeur à donner aux capacités commutées.

Exemple :

Soit à réaliser avec les dispositifs à commutation de capacités le filtre de Butterworth d'ordre 4 dont le schéma analogique est celui de la figure 8.12.*a*.

La procédure décrite au paragraphe précédent conduit au schéma de la figure 8.12.*b* pour une réalisation avec des intégrateurs, en supposant unitaires les résistances de terminaisons. Le schéma avec DCC est donné par la figure 8.12.*c*. Les coefficients a_i (i = 1, 2, 3, 4) qui définissent les rapports de capacités sont calculés par les relations (8.18).

Si le filtre a une fréquence f_c d'affaiblissement à 3 dB égale à 1 kHz les paramètres analogiques sont les suivants :

$$R1 = R2 = 1$$
$$C_2 = 121{,}8.10^{-6};\quad C_4 = 294{,}1.10^{-6}$$
$$L_3 = 294{,}1.10^{-6};\quad L_5 = 121{,}8.10^{-6}$$

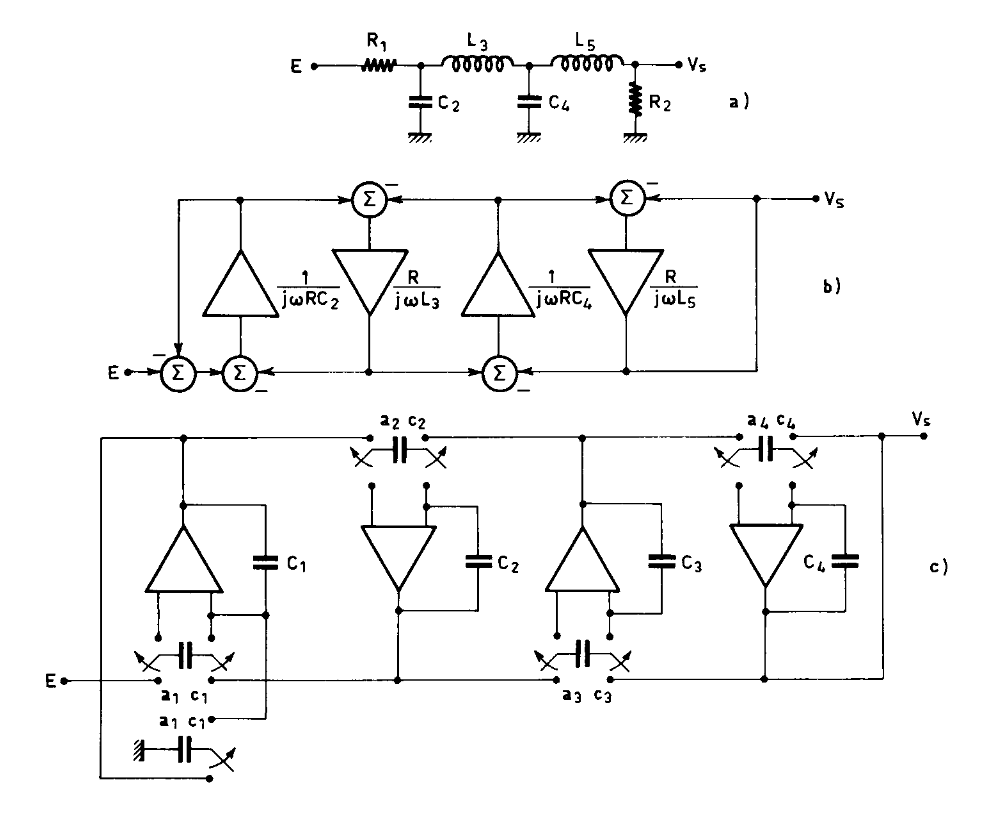


FIG. 8.12. *Filtre d'ordre 4 à commutation de capacités*

Avec une fréquence d'échantillonnage de 40 kHz, il vient :

$$a_1 = \frac{1}{4{,}87} = 0{,}205 = a_4$$

$$a_2 = \frac{1}{11{,}76} = 0{,}085 = a_3$$

Finalement, dans les dispositifs à commutation de capacités, la précision et la stabilité de la constante de temps d'un intégrateur dépendent de la fréquence d'échantillonnage fournie extérieurement et d'un rapport de capacités. Ces dispositifs permettent la réalisation de filtres très sélectifs sur une pastille de silicium, sous la forme d'un circuit intégré monolithique.

8.4 LES FILTRES D'ONDE

La première approche proposée pour la simulation numérique des filtres analogiques en échelle a été basée sur une simulation élément par élément, en considérant les variables d'onde. Des structures générales de filtres numériques condui-

sant à des coefficients représentables par un faible nombre de bits peuvent être ainsi élaborées. Elles sont désignées par filtres d'onde [6, 7].

Le filtre numérique est supposé offrir quatre accès à quatre suites de nombres comme indiqué sur la figure 8.13.

FIG. 8.13. *Accès du filtre d'onde*

L'onde incidente est constituée par les éléments de la suite à filtrer $a_1(n)$. L'onde transmise correspond aux éléments de la suite $b_2(n)$ filtrée par exemple par un passe-bas. L'onde réfléchie $b_1(n)$ correspond à la suite filtrée par le filtre complémentaire, par exemple passe-haut. On remarque que le dispositif réalise les deux filtres à la fois. La quatrième suite $a_2(n)$ est supposée nulle.

Comme aux paragraphes précédents la détermination du filtre numérique se fait à partir du filtre analogique en échelle, en deux étapes. D'abord les éléments numériques correspondant aux éléments analogiques sont définis. Aux bornes d'une inductance ou d'une capacité les ondes incidente A et réfléchie B s'écrivent :

$$\begin{aligned} A &= V + RI \\ B &= V - RI \end{aligned} \qquad (8.21)$$

où V et I sont la tension et le courant aux bornes, R une résistance de référence. Pour une inductance L on a :

$$V = LsI$$

Un équivalent numérique peut être obtenu très simplement en faisant appel à la transformation suivante, qui est la transformation homographique introduite au chapitre précédent, au facteur d'échelle $\frac{2}{T}$ près :

$$s = \frac{1 - Z^{-1}}{1 + Z^{-1}}$$

En prenant comme résistance de référence R = L, la relation suivante est établie entre les ondes incidente et réfléchie aux bornes du circuit numérique équivalent à l'inductance L :

$$B = -Z^{-1}A \qquad (8.22)$$

Ainsi l'équivalent numérique comporte un retard unitaire et une multiplication par – 1. Pour une capacité l'équivalent numérique est simplement un retard unitaire.

Dans cette procédure il apparaît que chaque élément possède une résistance de référence particulière et la deuxième phase de la détermination du filtre numérique consiste à interconnecter convenablement les éléments qui ont des résis-

tances de référence distinctes, ce qui nécessite des dispositifs d'adaptation. Pour mettre en série deux éléments il faut utiliser un adaptateur série dont le symbole est donné par la figure 8.14.*a*.

FIG. 8.14.*a*. *Connexion et symbole de l'adaptateur série*

Les relations entre les couples de variables d'onde (a_1, b_1), (a_2, b_2) et (a_3, b_3) sont obtenues en tenant compte des relations qui caractérisent cet adaptateur :

$$I_1 = I_2 = I_3$$
$$V_1 + V_2 + V_3 = 0 \tag{8.23}$$

FIG. 8.14.*b*. *Schéma de l'adaptateur série*

Le schéma de l'adaptateur série est donné par la figure 8.14.*b*. Les ondes réfléchies s'expriment par :

$$b_1 = a_1 - \beta_1(a_1 + a_2 + a_3) \quad \text{avec} \quad \beta_1 = \frac{2R_1}{R_1 + R_2 + R_3}$$

$$b_2 = a_2 - \beta_2(a_1 + a_2 + a_3) \quad \text{avec} \quad \beta_2 = \frac{2R_2}{R_1 + R_2 + R_3}$$

$$b_3 = -(a_1 + a_2 + a_3) - b_1 - b_2 \tag{8.24}$$

Pour la connexion parallèle, le symbole et le schéma sont donnés par la figure 8.15. Les relations qui caractérisent l'adaptateur correspondant sont les suivantes :

$$\begin{aligned} I_1 + I_2 + I_3 &= 0 \\ V_1 = V_2 &= V_3 \end{aligned} \quad (8.25)$$

FIG. 8.15. *Symbole et schéma de l'adaptateur parallèle*

Comme ci-dessus ces relations conduisent à déterminer les ondes réfléchies par les expressions :

$$b_k = a_0 - a_k \qquad \text{avec} \qquad k = 1, 2, 3.$$

et :

$$\begin{aligned} a_0 &= 2a_3 + \alpha_1(a_1 - a_3) + \alpha_2(a_2 - a_3) \\ \alpha_k &= \frac{2G_k}{G_1 + G_2 + G_3} \; ; \qquad G_k = \frac{1}{R_k} \end{aligned} \quad (8.26)$$

Lors des interconnexions d'adaptateurs entre eux, il peut apparaître des boucles sans retard, c'est-à-dire qu'à un accès, une onde réfléchie peut dépendre de l'onde incidente au même accès qui dépend elle-même de l'onde réfléchie à calculer. Il est possible d'ouvrir cette boucle en imposant à la résistance de référence à cet accès une valeur telle que l'onde réfléchie ne dépende pas de l'onde incidente. Cette condition est réalisée à l'accès 2 par exemple si, dans l'équation donnant b_2, le coefficient de a_2 est nul, ce qui pour l'adaptateur série implique : $R_2 = R_1 + R_3$.

Il vient alors :

$$\beta_1 = \frac{R_1}{R_1 + R_3}\ ; \quad \beta_2 = 1$$

Pour l'adaptateur parallèle : $G_2 = G_1 + G_3$ avec :

$$\alpha_1 = \frac{G_1}{G_1 + G_3} \quad \text{et} \quad \alpha_2 = 1$$

Ces adaptateurs sont dits ouverts à droite.

Tous les éléments nécessaires à la détermination du filtre sont ainsi disponibles.

Exemple :

Soit le filtre de la figure 8.16 avec les valeurs :

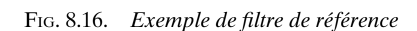
$$R_1 = R_2 = 1\ \Omega$$
$$C_1 = 0{,}1086\ \text{F}; \qquad C_2 = 0{,}1907\ \text{F}$$
$$L_1 = 0{,}2091\ \text{H}; \qquad L_2 = 0{,}0202\ \text{H}; \qquad L_3 = 0{,}09021\ \text{H}$$

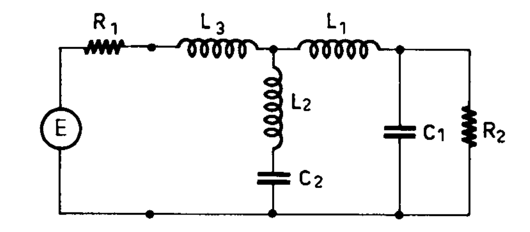
Ce filtre a 5 éléments, il faut 5 mémoires et 5 adaptateurs.

Le schéma du filtre est donné par la figure 8.17.

Les adaptateurs numérotés 1, 3 et 4 sont de type série ouverts à un accès et ont pour coefficients :

$$\beta_1 = 0{,}91725$$
$$\beta_3 = 0{,}99616$$
$$\beta_4 = 0{,}8120$$


FIG. 8.16. *Exemple de filtre de référence*



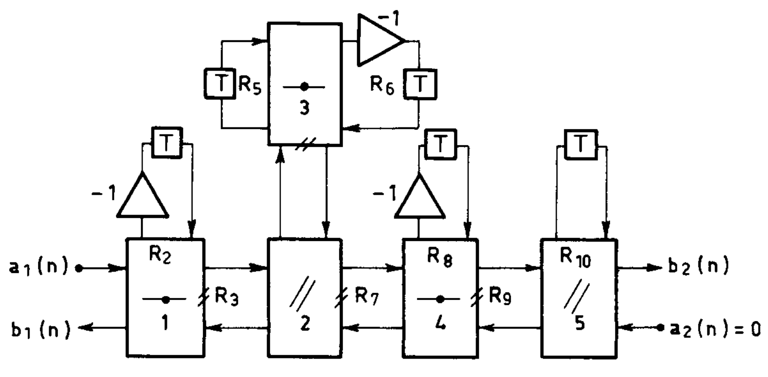


FIG. 8.17. *Schéma synoptique d'un filtre d'onde*

L'adaptateur 2 est du type parallèle ouvert à un accès; il a pour coefficient $\alpha_2 = 0{,}8284$. L'adaptateur 5 est de type parallèle général, avec deux coefficients qui sont : $\alpha_5 = 0{,}8956$ et $\alpha'_5 = 0{,}99618$.

FIG. 8.18. *Réponse en fréquence du filtre d'onde*

La fonction de transfert en Z correspondant à l'onde transmise est obtenue à partir des éléments donnés au paragraphe 8.1. Elle se met sous la forme suivante :

$$H(Z) = \frac{(Z+1)^2[Z^2 + 2Z(2\beta_3 - 1) + 1]}{D(Z)} \tag{8.27}$$

Elle présente un zéro double au point – 1 et deux zéros sur le cercle unité Z_1 et $\overline{Z}_1$ tels que :

$$Z_1 = -(2\beta_3 - 1) + j\sqrt{1-(2\beta_3-1)^2} \tag{8.28}$$

qui ne dépendent que du coefficient β_3.

La figure 8.18 représente la réponse en fréquence du filtre obtenu sans arrondi des coefficients et avec un arrondi à 6 bits des coefficients autres que β_3. Les propriétés de faible sensibilité en bande passante sont nettement mises en évidence.

Cet exemple montre que les coefficients peuvent être représentés par un faible nombre de bits comme dans les filtres en échelle simulée. L'arrondi du coefficient β_3 se traduit par un déplacement de la pointe d'affaiblissement infini, comme en structure cascade.

Une propriété intéressante des filtres d'onde est que les mêmes calculs et les mêmes circuits fournissent en même temps le filtre complémentaire. Le signal de sortie filtré correspond alors à la suite $b_1(n)$ de la figure 8.13. L'affaiblissement en fréquence de ce filtre est donnée à la figure 8.19.

FIG. 8.19. *Affaiblissement du filtre de réflexion*

Les filtres d'onde ont été obtenus par une analogie avec les lignes de transmission ; une analogie avec les tuyaux sonores a conduit aux filtres en treillis.

8.5 LES FILTRES EN TREILLIS

La structure de treillis apparaît dans les études d'analyse et de synthèse de la parole, pour la simulation du conduit vocal, et plus généralement dans les systèmes de prédiction linéaire. Elle permet de réaliser des filtres de type RIF et de type RII [8].

Soit la structure à M cellules représentée sur la figure 8.20.

FIG. 8.20. *Filtre en treillis de type RIF*

Les suites de sortie de la première cellule, $y_1(n)$ et $u_1(n)$ sont liées à la suite d'entrée $x(n)$ par les relations suivantes :

$$\begin{aligned} y_1(n) &= x(n) + k_1 x(n-1) \\ u_1(n) &= k_1 x(n) + x(n-1) \end{aligned} \tag{8.29}$$

De même les suites de sortie de la deuxième cellule, $y_2(n)$ et $u_2(n)$, sont liées à la suite d'entrée par :

$$\begin{aligned} y_2(n) &= x(n) + k_1(1+k_2)x(n-1) + k_2 x(n-2) \\ u_2(n) &= k_2 x(n) + k_1(1+k_2)x(n-1) + x(n-2) \end{aligned} \tag{8.30}$$

Par itération il apparaît que les suites de sortie du filtre, $y_M(n)$ et $u_M(n)$, sont liées à la suite $x(n)$ par les relations suivantes, qui correspondent à un filtrage de type RIF :

$$y_M(n) = \sum_{i=0}^{M} a_i x(n-i) \tag{8.31}$$

$$u_M(n) = \sum_{i=0}^{M} a_{M-i} x(n-i) \tag{8.32}$$

Les deux filtres RIF ainsi obtenus ont les mêmes coefficients mais dans l'ordre inverse ; leurs fonctions de transfert en Z, $H_M(Z)$ et $U_M(Z)$, sont des polynômes images. Il vient :

$$H_M(Z) = \sum_{i=0}^{M} a_i Z^{-i}$$

$$U_M(Z) = \sum_{i=0}^{M} a_{M-i} Z^{-i} = Z^{-M} H_M(Z^{-1}) \tag{8.33}$$

Pour déterminer les coefficients k_i du filtre en treillis à partir des coefficients a_i $(1 \leqslant i \leqslant M)$, il faut procéder par itérations.

D'abord le coefficient a_0 est supposé égal à l'unité. Ensuite il est aisé de vérifier d'après les relations données précédemment, et aussi directement sur la figure 8.20, que l'on a :

$$k_M = a_M$$

Cette remarque est à la base du calcul. En désignant par $H_m(Z)$ et $U_m(Z)$ $(1 \leqslant m \leqslant M)$, les fonctions de transfert correspondant aux sorties de la cellule de rang m, on peut écrire la relation matricielle suivante :

$$\begin{bmatrix} H_m(Z) \\ U_m(Z) \end{bmatrix} = \begin{bmatrix} 1 & k_m Z^{-1} \\ k_m & Z^{-1} \end{bmatrix} \begin{bmatrix} H_{m-1}(Z) \\ U_{m-1}(Z) \end{bmatrix}$$

Cette relation peut aussi s'écrire, en supposant $k_m \neq 1$:

$$\begin{bmatrix} H_{m-1}(Z) \\ U_{m-1}(Z) \end{bmatrix} = \frac{1}{1-k_m^2} \begin{bmatrix} 1 & -k_m \\ -k_m Z & Z \end{bmatrix} \begin{bmatrix} H_m(Z) \\ U_m(Z) \end{bmatrix} \tag{8.34}$$

Ainsi les polynômes $H_{m-1}(Z)$ et $U_{m-1}(Z)$ sont des polynômes images de degré $m-1$ dont les coefficients $a_{i(m-1)}$ $(1 \leqslant i \leqslant m-1)$ se calculent à partir des coefficients a_{im} $(1 \leqslant i \leqslant m)$ des polynômes $H_m(Z)$ et $U_m(Z)$.

Dans ces conditions il vient :

$$a_{mm} = k_m ; \quad a_{(m-1)(m-1)} = k_{m-1}$$

$$a_{(m-1)(m-1)} = \frac{1}{1 - a_{mm}^2} \left[a_{(m-1)m} - a_{mm} a_{1m} \right] \tag{8.35}$$

Les coefficients k_m $(1 \leqslant m \leqslant \mathrm{M})$ sont ainsi calculés en M itérations.

Exemple :

Soit la fonction de transfert $H_3(Z)$ telle que :

$$H_3(Z) = 1 - 1{,}990\ Z^{-1} + 1{,}572\ Z^{-2} - 0{,}4583\ Z^{-3}$$

Il vient :

$$k_3 = -\,0{,}4583$$

D'après la relation (8.34), on peut écrire :

$$H_2(Z) = 1 - 1{,}607\ Z^{-1} + 0{,}8355\ Z^{-2}$$

D'où :

$$k_2 = 0{,}8355$$

Par application de la relation (8.35), il vient :

$$k_1 = -\,0{,}8756 \quad \text{et} \quad H_1(Z) = 1 - 0{,}8756\ Z^{-1}$$

La réalisation de filtres de type RII purement récursif conduit à une structure duale, représentée sur la figure 8.21.

FIG. 8.21. *Filtre en treillis de type RII purement récursif*

Les suites $x_1(n)$, $u_1(n)$ et $y(n)$ sont liées par les relations suivantes :

$$y(n) = x_1(n) - k_1 y(n-1)$$
$$u_1(n) = k_1\, y(n) + y(n-1)$$

De même les suites $x_2(n)$, $x_1(n)$, $u_1(n)$ et $u_2(n)$ sont liées par les relations :

$$x_1(n) = x_2(n) - k_2 u_1(n-1)$$
$$u_2(n) = k_2 x_1(n) + u_1(n-1)$$

Il en résulte entre l'entrée $x_2(n)$ et la sortie $y(n)$ la fonction de transfert $H_2(Z)$ telle que :

$$H_2(Z) = \frac{1}{1 + k_1(1 + k_2)Z^{-1} + k_2 Z^{-2}}$$

De même entre $u_2(n)$ et $y(n)$ apparaît la fonction de transfert $U_2(Z)$ telle que :

$$U_2(Z) = k_2 + k_1(1 + k_2)Z^{-1} + Z^{-2}$$

Par itération il apparaît que les suites $x_M(n)$ et $y(n)$ d'une part, $u_M(n)$ et $y(n)$ d'autre part sont liées par les relations :

$$y(n) = x_M(n) - \sum_{i=1}^{M} b_i y(n-i) \qquad (8.36)$$

$$u_M(n) = \sum_{i=0}^{M-1} b_{M-i} y(n-i) + y(n-M) \qquad (8.37)$$

Il en résulte les fonctions de transfert $H_M(Z)$ et $U_M(Z)$ telles que :

$$H_M(Z) = \frac{1}{1 + \sum_{i=1}^{M} b_i Z^{-i}} = \frac{1}{D_M(Z)}$$

$$U_M(Z) = \sum_{i=0}^{M-1} b_{M-i} Z^{-i} + Z^{-M} = Z^{-M} D_M(Z^{-1}) \qquad (8.38)$$

Pour calculer les coefficients k_i $(1 \leqslant i \leqslant M)$ du filtre en treillis à partir des coefficients b_i du filtre RII il faut procéder par itérations en remarquant que l'on a :

$$k_M = b_M$$

En désignant par $H_m(Z)$ et $U_m(Z)$ les fonctions relatives à l'ensemble de m cellules $(1 \leqslant m \leqslant M)$, il est possible, à partir des équations de définition :

$$x_{m-1}(n) = x_m(n) - k_m u_{m-1}(n-1)$$
$$u_m(n) = k_m x_{m-1}(n) + u_{m-1}(n-1)$$

de faire apparaître la relation matricielle suivante :

$$\begin{bmatrix} D_m(Z) \\ U_m(Z) \end{bmatrix} = \begin{bmatrix} 1 & k_m Z^{-1} \\ k_m & Z^{-1} \end{bmatrix} \begin{bmatrix} D_{m-1}(Z) \\ U_{m-1}(Z) \end{bmatrix}$$

Comme dans le cas du type de filtre RIF, cette relation matricielle s'écrit aussi, pour $k_m \neq 1$:

$$\begin{bmatrix} D_{m-1}(Z) \\ U_{m-1}(Z) \end{bmatrix} = \frac{1}{1 - k_m^2} \begin{bmatrix} 1 & -k_m \\ -k_m Z & Z \end{bmatrix} \begin{bmatrix} D_m(Z) \\ U_m(Z) \end{bmatrix} \qquad (8.39)$$

Cette expression permet, comme précédemment pour les filtres RIF, de calculer, à partir du polynôme $D_M(Z)$ tel que :

$$D_M(Z) = 1 + \sum_{i=1}^{M} b_i Z^{-i}$$

les coefficients k_i $(1 \leqslant i \leqslant M)$ du filtre RII en treillis, en M itérations.

Les structures en treillis données par les figures 8.20 et 8.21 sont canoniques pour les mémoires de données, mais pas pour les multiplications. Elles peuvent être rendues canoniques, par exemple en utilisant pour le type RII la cellule à une multiplication représentée sur la figure 8.22. Par contre, il faut alors une addition de plus.

FIG. 8.22. *Cellule de filtre en treillis à une seule multiplication*

Les équations de cette cellule à l'ordre 1 sont les suivantes :

$$(1 + k)x_1(n) = y(n) + ky(n-1)$$
$$(1 + k)u_1(n) = ky(n) + y(n-1)$$

Au facteur $(1 + k)$ près elles sont bien équivalentes à celles du treillis à deux multiplieurs.

Contrairement aux structures décrites dans les paragraphes précédents, les filtres en treillis ne présentent pas d'avantages particuliers pour le nombre de bits nécessaires à la représentation des coefficients. Cependant une propriété intéressante en pratique est la suivante : une condition nécessaire et suffisante pour que le filtre RII ait tous ses pôles à l'intérieur du cercle unité et donc soit stable, est que les coefficients soient en module inférieurs à l'unité.

$$|k_i| < 1; \qquad 1 \leqslant i \leqslant M$$

Cette propriété est évidente pour k_1 sur la figure 8.21, si l'on isole la cellule correspondante; elle s'étend aux autres coefficients en considérant les sous-ensembles du circuit, et en raisonnant par récurrence.

Il en résulte un contrôle de stabilité très simple à réaliser et particulièrement utile dans les systèmes où les valeurs des coefficients évoluent en permanence, comme les filtres adaptatifs.

Les structures de treillis considérées ci-dessus sont soit non récursive soit purement récursive. Il faut remarquer que la structure purement récursive peut être complétée pour faire un filtre général, il suffit de former une sommation pondérée des variables $u_m(n)$. En effet, la relation :

$$v_M(n) = \gamma_0\, y(n) + \sum_{m=1}^{M} \gamma_m u_m(n)$$

définit un filtrage de type RIF sur le signal $y(n)$, en raison des relations (8.37).
Les coefficients b_i $(1 \leqslant i \leqslant M)$ étant fixés, les coefficients γ_i peuvent être calculés pour obtenir un numérateur quelconque pour le filtre général.

Il est intéressant d'observer également que la structure purement récursive comporte la fonction de déphaseur pur. En effet, les équations (8.37) et (8.38) permettent d'écrire :

$$H_D(Z) = \frac{U_M(Z)}{X(Z)} = \frac{b_M + b_{M-1}Z^{-1} + \ldots + Z^{-M}}{1 + b_1 Z^{-1} + \ldots + b_M Z^{-M}}$$

Cette expression montre que, comme indiqué au paragraphe (6.3), le signal $u_M(n)$ est la sortie d'un déphaseur pur dont $x(n)$ est l'entrée. La fonction de transfert $H_D(Z)$ s'exprime directement en fonction des coefficients du treillis par une fraction continue :

$$H_D(Z) = k_M + \cfrac{(1-k_M^2)Z^{-1}}{k_M Z^{-1} + \cfrac{1}{k_{M-1}Z^{-1} + \cfrac{(1-k_{M-1}^2)Z^{-1}}{\ddots + \cfrac{1}{k_1 + \cfrac{(1-k_1^2)Z^{-1}}{k_1 Z^{-1} + 1}}}}}$$

Cette remarque peut être utilisée pour calculer directement les pôles du filtre en treillis [9].

Une application intéressante des résultats ci-dessus est le filtre à encoche introduit au paragraphe 6.3. La sortie du filtre à encoche $y_E(n)$ est obtenue simplement en ajoutant un additionneur au circuit de la figure 8.21, pour effectuer l'opération :

$$y_E(n) = x_M(n) + u_M(n) \qquad (8.40)$$

A l'ordre 2, la fonction de transfert du déphaseur s'écrit :

$$H_D(Z) = \frac{k_2 + k_1(1+k_2)Z^{-1} + Z^{-2}}{1 + k_1(1+k_2)Z^{-1} + k_2 Z^{-2}}$$

Une propriété utile de cette approche est que la fréquence ω_0 et la bande d'affaiblissement à 3 dB, B_{3E}, peuvent être ajustées indépendamment [10]. Ce découplage provient des relations suivantes :

$$k_1 \simeq -\cos \omega_0$$

$$k_2 \simeq \frac{1 - \mathrm{tg}\, \pi B_{3E}}{1 + \mathrm{tg}\, \pi B_{3E}} = (1 - \varepsilon)^2 \tag{8.41}$$

Si l'on fait une soustraction dans l'équation (8.40) au lieu d'une addition, c'est le filtre complémentaire que l'on obtient.

8.6 ÉLÉMENTS DE COMPARAISON

Après la présentation des diverses structures de filtres numériques il est utile de faire une récapitulation de leurs propriétés. La référence [11] présente une analyse comparative détaillée.

La structure la plus facile à obtenir est la structure cascade, puisque les coefficients correspondent à une simple factorisation de la fonction de transfert en Z. Elle conduit au minimum de multiplications, d'additions et de mémoires. Par contre la représentation des coefficients peut nécessiter un nombre de bits important.

Le choix entre les structures cascade et treillis ne se présente généralement pas pour les filtres à coefficients fixes car la structure en treillis correspond à des utilisations particulières.

Les structures tirées de la simulation des réseaux analogiques en échelle, les filtres en échelle simulée et les filtres d'onde, offrent une représentation des coefficients qui peut se limiter à quelques bits même pour des filtres très sélectifs ; il s'en suit que la suppression des multiplieurs est envisageable ; comme ce sont généralement les circuits les plus complexes, le gain en matériel est appréciable. Cependant il faut tenir compte d'un certain nombre de complications. D'abord le nombre d'additions est augmenté. Ensuite, l'enchaînement des opérations se complique car, par exemple dans les filtres d'onde, il faut faire les calculs correspondant aux ondes incidentes et réfléchies à l'intérieur du filtre et exécuter ainsi un aller et retour entre l'entrée et la sortie. Ces opérations exigent des mémorisations complémentaires pour des résultats intermédiaires. D'autre part le multiplexage des opérations entre plusieurs filtres, qui est un avantage important du traitement numérique, est rendu difficile. Finalement une évaluation complète est nécessaire avant de retenir ce type de structure.

BIBLIOGRAPHIE

[1] M. Feldmann – *Théorie des Réseaux et Systèmes Linéaires.* Ed. Eyrolles, Collection CNET-ENST, Paris, 1981.

[2] J. Neirynk et Ph. Van Bastelaer – La synthèse des filtres par factorisation de la matrice de transfert. *Revue MBLE*, Belgique, Vol. 10, N° 1, 1967.

[3] L. T. Bruton – Low Sensitivity Digital Ladder Filters. *IEEE Trans,* Vol. CAS 22, N° 3, Mars 1975.

[4] B. Hosticka, R. Brodersen and P. Gray – MOS Sampled Data Recursive Filters Using Switched Capacitor Integrators. *IEEE Journal Solid-State Circuits*, Vol. SC-12, Dec. 1977.

[5] R. Brodersen, P. Gray and D. Hodges – MOS Switched-Capacitor Filters. *Proceedings of the IEEE,* Vol. 67, N° 1, January 1979.

[6] A. Fettweis – Digital Filter Structures Related to Classical Filter Networks. *Archiv. Elek. Ubertragung*, Vol. 25, Feb. 1971.

[7] J. R. Boite et H. Leich – Synthèse des filtres numériques simulant les filtres à inductances et capacités. *Annales des Télécom.* Tome 31, N° 3-4, Mars 1976.

[8] S. K. Mitra, P. S. Kamat and D. C. Huey – Cascaded Lattice Realization of Digital Filters. *Circuit Theory and Applications,* Vol. 5, 1977.

[9] W. B. Jones and A. O. Steinhardt – «Finding the Poles of the lattice filter», *IEEE Trans.,* Vol. ASSP-33, N° 4, Oct. 1985, pp. 1328-1331.

[10] T. Saramaki, T. H. Yu and S. K. Mitra – Very low Sensitivity Realization of IIR Digital Filters Using a Cascade of Complex All-pass Structures, *IEEE Trans.,* CAS-34, August 1987, pp. 876-886.

[11] R. E. Crochière and A. V. Oppenheim – Analysis of Linear Digital Networks. *Proceedings of the IEEE,* Vol. 63, N° 4, April 1975.

EXERCICES

1 Écrire pour les quadripôles élémentaires donnés à la figure 8.2, les matrices d'impédance et de répartition. Donner les matrices d'impédance, de répartition et de transfert quand les éléments sont des circuits résonants LC.

2 Soit le filtre de Butterworth d'ordre 4 donné à la figure 8.12.*a*. Tracer le graphe de fluence correspondant. Donner le schéma du filtre numérique en échelle simulée et calculer les coefficients en partant des valeurs données pour les éléments analogiques, avec une fréquence d'échantillonnage de 40 kHz. Examiner la modification de fonction de transfert en bande passante apportée par une réduction de la fréquence d'échantillonnage de 40 kHz à 10 kHz.

3 Tracer le graphe de fluence du filtre de la figure 8.16, qui comprend un circuit résonant dans une branche. Donner un schéma de réalisation avec dispositifs à commutation de capacités et calculer les rapports de capacités pour chaque intégrateur à partir des éléments analogiques en supposant une fréquence d'échantillonnage de 20 kHz. Comparer la réponse en fréquence obtenue à celle du filtre d'onde.

4 Déterminer les éléments du filtre d'onde correspondant au filtre passe-bas de Tchebycheff d'ordre 7 dont les éléments analogiques sont donnés au paragraphe 8.2. La fréquence d'échantillonnage est prise égale à 10 kHz. Donner le schéma du filtre numérique en échelle simulée correspondant.

Comparer les quantités d'opérations à faire dans chaque réalisation. Quelle est la plus avantageuse ?

Comparer les réponses en fréquence, quand les coefficients sont représentés par 5 bits.

5 Calculer la réponse en fréquence du filtre en treillis donné en exemple au paragraphe 8.5. Comment évolue cette réponse quand les paramètres sont représentés par 5 bits. Établir le schéma du filtre avec des cellules à une seule multiplication. Comment doit-on modifier le schéma pour obtenir le filtre de fonction de transfert en Z inverse.

Chapitre 9

Signaux complexes Filtres de quadrature Interpolateurs

Les signaux complexes, sous la forme de suites dont les éléments sont des nombres complexes, sont d'une utilisation courante en traitement numérique du signal. De telles suites ont été considérées par exemple au chapitre qui traite de la Transformation de Fourier Discrète. Dans le présent chapitre, une catégorie particulière de signaux complexes va être étudiée, qui offre des propriétés intéressantes, celles des signaux analytiques. Ces signaux interviennent principalement dans les processus de modulation et de multiplexage. Les propriétés des transformées de Fourier de suites réelles et causales vont être examinées d'abord [1, 2, 3].

9.1 TRANSFORMÉE DE FOURIER D'UNE SUITE RÉELLE ET CAUSALE

Soit une suite d'éléments $x(n)$ dont la transformée en Z s'écrit :

$$X(Z) = \sum_{n=-\infty}^{\infty} x(n) Z^{-n}$$

La transformée de Fourier de cette suite est obtenue en remplaçant Z par $e^{j2\pi f}$ dans $X(Z)$:

$$X(f) = \sum_{n=-\infty}^{\infty} x(n) e^{-j2\pi nf}$$

Si les éléments $x(n)$ sont des nombres réels, on a :

$$X(-f) = \overline{X(f)} \tag{9.1}$$

Les valeurs de $X(f)$ aux fréquences négatives sont complexes conjuguées des valeurs aux fréquences positives.

Une condition supplémentaire peut être imposée à la suite $x(n)$, d'être causale. Les conséquences sur $X(f)$ vont être examinées.

La fonction $X(f)$ peut être séparée en parties réelle et imaginaire :

$$X(f) = X_R(f) + jX_I(f) \tag{9.2}$$

Si la suite $x(n)$ est réelle, d'après la relation (9.1), la fonction $X_R(f)$ étant paire est la transformée de Fourier d'une suite paire $x_p(n)$ et la fonction $X_I(f)$ est la transformée de Fourier d'une suite impaire $x_i(n)$, telles que :

$$\begin{aligned} x_p(n) &= x_p(-n) \\ x_i(n) &= -x_i(-n) \\ x(n) &= x_p(n) + x_i(n) \end{aligned} \tag{9.3}$$

Il vient dans ces conditions :

$$X_R(f) = x_p(0) + 2\sum_{n=1}^{\infty} x_p(n)\cos(2\pi nf) \tag{9.4}$$

$$X_I(f) = -2\sum_{n=1}^{\infty} x_i(n)\sin(2\pi nf) \tag{9.5}$$

Si la suite $x(n)$ est causale, c'est-à-dire que :

$$x(n) = 0 \quad \text{pour} \quad n < 0$$

on a les relations (fig. 9.1) :

$$x_i(n) = x_p(n) = \frac{1}{2}x(n) \quad \text{pour} \quad n \geqslant 1$$

$$x_p(0) = x(0)$$

et il vient :

$$X_R(f) - x(0) = \sum_{n=1}^{\infty} x(n)\cos(2\pi nf) \tag{9.6}$$

$$X_I(f) = -\sum_{n=1}^{\infty} x(n)\sin(2\pi nf) \tag{9.7}$$

FIG. 9.1. *Décomposition d'une suite causale en parties paire et impaire*

Il apparaît que ces deux fonctions sont liées. Pour passer de l'une à l'autre il suffit de changer cos $(2\pi nf)$ en – sin $(2\pi nf)$ ou inversement ; une telle opération est appelée une quadrature, elle va être exprimée analytiquement.

Par définition une suite causale est une suite qui satisfait l'égalité :

$$x(n) = x(n) \, . \, \mathrm{Y}(n)$$

où la suite $\mathrm{Y}(n)$ est telle que :

$$\begin{aligned} \mathrm{Y}(n) &= 0 \quad \text{pour} \quad n < 0 \\ \mathrm{Y}(n) &= 1 \quad \text{pour} \quad n \geqslant 0 \end{aligned}$$

Cette suite est un échantillonnage de la fonction échelon unité $\mathrm{Y}(t)$ qui possède, au sens des distributions, une transformée de Fourier FY donnée par [2] :

$$\mathrm{FY} = \frac{1}{j2\pi} \, vp\left(\frac{1}{f}\right) + \frac{1}{2} \, \delta(f) \tag{9.8}$$

où $vp\left(\frac{1}{f}\right)$ est la distribution définie par l'expression :

$$\langle vp\left(\frac{1}{f}\right), \varphi \rangle = \mathrm{VP}\int_{-\infty}^{\infty} \frac{\varphi(f)}{f} \, df \tag{9.9}$$

la valeur principale de l'intégrale au sens de Cauchy étant elle-même définie par :

$$\mathrm{VP}\int_{-\infty}^{\infty} \frac{\delta(f)}{f} \, df = \fint_{-\infty}^{\infty} \frac{\varphi(f)}{f} \, df = \lim_{\varepsilon \to 0} \left[\int_{-\infty}^{-\varepsilon} \frac{\varphi(f)}{f} \, df + \int_{\varepsilon}^{\infty} \frac{\varphi(f)}{f} \, df \right]$$

L'échantillonnage introduisant une périodicité du spectre, on démontre, en introduisant la transformation bilinéaire et avec la relation (7.10), que la transformée de Fourier de la suite $\mathrm{Y}(n)$ telle que :

$$\begin{aligned} \mathrm{Y}(n) &= 0 \quad \text{pour} \quad n < 0 \\ \mathrm{Y}(0) &= 1/2 \\ \mathrm{Y}(n) &= 1 \quad \text{pour} \quad n > 0 \end{aligned}$$

est la distribution FY_n qui s'écrit :

$$\mathrm{FY}_n = \frac{1}{2j} \, vp[\mathrm{cotg}\, \pi f] + \frac{1}{2} \, \delta(f) \quad \text{pour} \quad -\frac{1}{2} \leqslant f \leqslant \frac{1}{2} \tag{9.10}$$

Au produit de deux suites correspond le produit de convolution des transformées de Fourier. Il vient :

$$\mathrm{X}(f) = \left[\frac{1}{2j} \, vp[\mathrm{cotg}\, \pi f] + \frac{1}{2} \, \delta(f) \right] * \mathrm{X}(f) + \frac{1}{2} \, x(0)$$

En séparant les parties réelles et imaginaires, on obtient :

$$\mathrm{X_R}(f) + j\mathrm{X_I}(f) = vp[\mathrm{cotg}\, \pi f] * [\mathrm{X_I}(f) - j\mathrm{X_R}(f)] + x(0)$$

Les relations qui lient les parties réelle et imaginaire de $X(f)$ s'expriment par :

$$X_R(f) = x(0) + \fint_{-\frac{1}{2}}^{\frac{1}{2}} X_I(f') \operatorname{cotg}[\pi(f-f')]\, df' \qquad (9.11)$$

$$X_I(f) = -\fint_{-\frac{1}{2}}^{\frac{1}{2}} X_R(f') \operatorname{cotg}[\pi(f-f')]\, df' \qquad (9.12)$$

ou encore sous une forme différente sans introduire les valeurs principales de Cauchy :

$$X_R(f) = x(0) - \int_{-\frac{1}{2}}^{\frac{1}{2}} [X_I(f) - X_I(f')] \operatorname{cotg}[\pi(f-f')]\, df' \qquad (9.13)$$

$$X_I(f) = \int_{-\frac{1}{2}}^{\frac{1}{2}} [X_R(f) - X_R(f')] \operatorname{cotg}[\pi(f-f')]\, df' \qquad (9.14)$$

Les parties réelle et imaginaire de la transformée de Fourier d'une suite causale sont liées par les relations (9.11) et (9.12) qui correspondent à la transformation de Hilbert pour les signaux continus.

Exemple :

Soit la suite $U_k(n)$ telle que :

$$U_k(k) = 1$$
$$U_k(n) = 0 \quad \text{pour} \quad n \neq k$$
$$X_R(f) = \cos(2\pi k f); \quad X_I(f) = -\sin(2\pi k f)$$

On vérifie directement que :

$$\fint_{-\frac{1}{2}}^{\frac{1}{2}} \cos(2\pi k f') \operatorname{cotg}[\pi(f-f')]\, df' = \fint_{-\frac{1}{2}}^{\frac{1}{2}} \cos[2\pi k(f-f')] \operatorname{cotg}(\pi f')\, df' = \sin(2\pi k f)$$

$$\fint_{-\frac{1}{2}}^{\frac{1}{2}} \sin(2\pi k f') \operatorname{cotg}[\pi(f-f')]\, df' = \fint_{-\frac{1}{2}}^{\frac{1}{2}} \sin[2\pi k(f-f')] \operatorname{cotg}(\pi f')\, df' = -\cos(2\pi k f)$$

D'autre part d'après la relation de Parseval on peut écrire :

$$\int_0^{\frac{1}{2}} X_R^2(f)\, df = \int_0^{\frac{1}{2}} X_I^2(f)\, df$$

La partie réelle et la partie imaginaire de $X(f)$ ont la même puissance.

9.2 SIGNAL ANALYTIQUE

Les signaux analytiques correspondent aux signaux causaux quand on échange temps et fréquence. Leur spectre ne contient pas de composantes aux fréquences négatives et leur dénomination provient du fait qu'ils constituent la restriction à

l'axe réel d'une fonction de variable complexe analytique c'est-à-dire développable en série entière dans une région contenant cet axe.

Les propriétés des signaux analytiques se déduisent de celles des signaux causaux en échangeant temps et fréquence.

Soit le signal : $x(t) = x_R(t) + jx_I(t)$ tel que :

$$X(f) = 0 \quad \text{pour} \quad f < 0$$

Les fonctions $x_R(t)$ et $x_I(t)$ sont transformées de Hilbert l'une de l'autre :

$$x_R(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x_I(t')}{t-t'}\, dt' \tag{9.16}$$

$$x_I(t) = -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x_R(t')}{t-t'}\, dt' \tag{9.17}$$

D'autre part la transformée de Fourier de la fonction réelle :

$$x_R(t) = \frac{1}{2}\, [x(t) + \bar{x}(t)] \tag{9.18}$$

est la fonction $X_R(f)$ telle que :

$$X_R(f) = \frac{1}{2}\, [X(f) + \overline{X}(-f)] \tag{9.19}$$

c'est-à-dire que $X_R(f) = \frac{1}{2} X(f)$ pour les fréquences positives et $X_R(f) = \frac{1}{2}\overline{X}(-f)$ pour les fréquences négatives.

De même :

$$X_I(f) = -j\,\frac{1}{2}\,[X(f) - \overline{X}(-f)] \tag{9.20}$$

La figure 9.2 illustre la décomposition du spectre d'un signal en parties réelles et imaginaires.

Exemple :

$$x(t) = e^{j\omega t}; \quad x_R(t) = \cos \omega t = \frac{1}{2}\,[e^{j\omega t} + e^{-j\omega t}]$$

$$x_I(t) = \sin \omega t = -j\,\frac{1}{2}\,[e^{j\omega t} - e^{-j\omega t}]$$

Il apparaît finalement entre $X_R(f)$ et $X_I(f)$ les relations suivantes :

$$X_I(f) = -jX_R(f) \quad \text{pour} \quad f > 0$$
$$X_I(f) = jX_R(f) \quad \text{pour} \quad f < 0$$

C'est-à-dire que $X_I(f)$ est obtenu à partir de $X_R(f)$ par une rotation égale à $\frac{\pi}{2}$ des composantes. La transformation de Hilbert consiste en une mise en quadrature des

composantes du signal; c'est une opération de filtrage avec la réponse en fréquence $Q(f)$ représentée sur la figure 9.3.

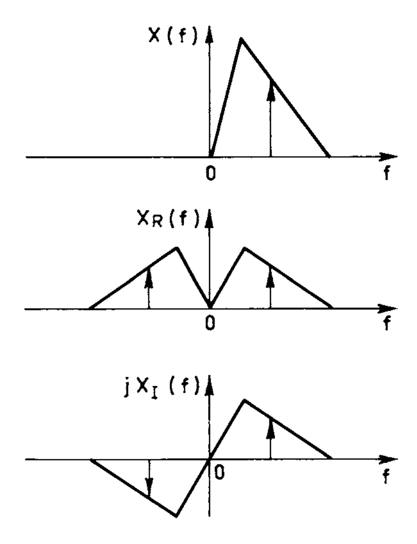


FIG. 9.2. *Spectre d'un signal analytique*

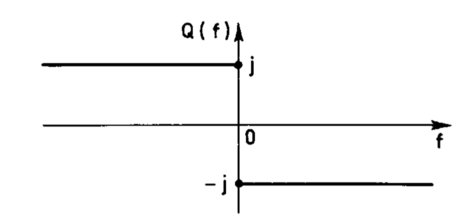


FIG. 9.3. *Réponse en fréquence du filtre de quadrature*

Exemple :

$$x_R(t) = \int_0^{\infty} [A(f)\cos(2\pi ft) - B(f)\sin(2\pi ft)]\,df \qquad (9.21)$$

$$x_I(t) = \int_0^{\infty} [A(f)\sin(2\pi ft) + B(f)\cos(2\pi ft)]\,df \qquad (9.22)$$

Les propriétés des signaux analytiques continus peuvent se transposer aux signaux discrets moyennant certaines adaptations.

Un signal discret a une transformée de Fourier périodique. Un signal discret analytique $x(n)$ déduit d'un signal réel, est un signal discret dont la transformée de Fourier $X_n(f)$, qui a la période $f_e = 1$, s'annule pour $-\frac{1}{2} \leqslant f < 0$ (fig. 9.4).

Si un signal discret $x(n)$ est obtenu par échantillonnage d'un signal continu analytique $x(t)$ à la fréquence $f_e = 1$, il convient de remarquer que la restitution de

signal continu à partir des valeurs discrètes est obtenue par un filtre de restitution qui ne conserve que les composantes du signal comprises dans la bande $(0, f_e)$ comme le montre la figure 9.4. La formule de restitution correspondant à l'expression (1.57), s'écrit :

$$x(t) = \sum_{n=-\infty}^{\infty} x(n) \frac{\sin[\pi(t-n)]}{\pi(t-n)} e^{j\pi(t-n)} \qquad (9.23)$$

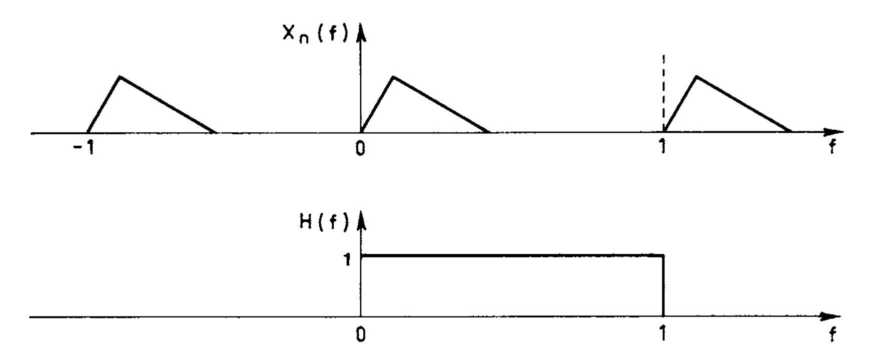


FIG. 9.4. *Spectre d'un signal analytique discret et filtre d'interpolation*

Par suite l'échantillonnage n'apporte pas de dégradation au signal analytique $x(t)$ si son spectre ne contient pas de composantes aux fréquences supérieures ou égale à f_e. D'où le théorème de l'échantillonnage pour un signal analytique :

Un signal analytique qui ne contient pas de composantes à des fréquences supérieures ou égale à f_m est entièrement déterminé par la suite de ses valeurs prélevées à des instants espacés de $T = \frac{1}{f_m}$.

La suite $x(n)$ se décompose en une suite réelle $x_R(n)$ et une suite imaginaire $x_I(n)$, telle que :

$$x(n) = x_R(n) + jx_I(n)$$

Les transformées de Fourier correspondantes $X_{nR}(f)$ et $X_{nI}(f)$ sont obtenues à partir de la transformée de Fourier $X_n(f)$ par les relations (9.19) et (9.20) données précédemment.

$$X_{nI}(f) = -jX_{nR}(f) \quad \text{pour} \quad 0 < f < \frac{1}{2}$$

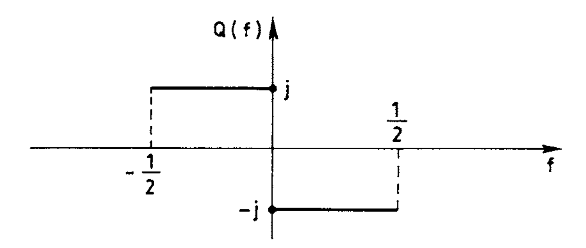
$$X_{nI}(f) = jX_{nR}(f) \quad \text{pour} \quad -\frac{1}{2} < f < 0$$

Les relations entre les suites $x_R(n)$ et $x_I(n)$ sont obtenues en considérant le filtre de quadrature dont la réponse en fréquence est donnée par la figure 9.5. La réponse impulsionnelle de ce filtre est la suite $h(n)$, telle que :

$$h(n) = \int_{-\frac{1}{2}}^{0} j \, . \, e^{j2\pi nf} \, df + \int_{0}^{\frac{1}{2}} (-j) \, . \, e^{j2\pi nf} \, df$$

$$h(n) = \frac{2}{\pi n} \sin^2\left(\frac{n\pi}{2}\right) \quad \text{pour} \quad n \neq 0 \tag{9.24}$$

$$h(0) = 0$$



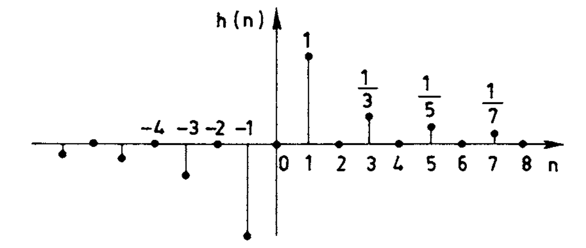


FIG. 9.5. *Réponses du filtre de quadrature*

En appliquant à ce filtre la suite $x_R(n)$, on obtient la suite $x_I(n)$, d'où :

$$x_I(n) = \frac{2}{\pi} \sum_{\substack{m=-\infty \\ m \neq n}}^{\infty} x_R(n-m) \, \frac{\sin^2\left(\pi \frac{m}{2}\right)}{m} \tag{9.25}$$

de même :

$$x_R(n) = -\frac{2}{\pi} \sum_{\substack{m=-\infty \\ m \neq n}}^{\infty} x_I(n-m) \, \frac{\sin^2\left(\pi \frac{m}{2}\right)}{m} \tag{9.26}$$

Les suites $x_R(n)$ et $x_I(n)$ sont liées par la transformation dite de Hilbert discrète [4].

L'examen des éléments de la suite $h(n)$ amène plusieurs remarques. D'abord le fait qu'un élément sur deux soit nul entraîne que si la suite $x_R(n)$ a également un

élément sur deux nul, il en est de même de la suite $x_I(n)$ et les deux suites $x_R(n)$ et $x_I(n)$ sont entrelacées. Un exemple sera donné ultérieurement.

D'autre part la réponse impulsionnelle du filtre de quadrature correspond à un cas de filtre RIF à phase linéaire mentionné au paragraphe 5.2. En effet sa réponse en fréquence s'écrit :

$$Q(f) = -j \cdot 2 \sum_{n=1}^{\infty} h(n) \sin(2\pi n f) \tag{9.27}$$

Pour la réalisation il faut limiter le nombre de coefficients.

9.3 CALCUL DES COEFFICIENTS D'UN FILTRE DE QUADRATURE RIF

Un filtre de quadrature réalisable est obtenu simplement en limitant le nombre de termes sur lesquels porte la sommation (9.27). La réponse en fréquence s'écarte alors de la réponse idéale. En pratique le filtre est spécifié par un gabarit donnant l'ondulation tolérée δ dans une bande de fréquence (f_1, f_2) comme le montre la figure 9.6. Pour aboutir à un filtre RIF satisfaisant on peut partir d'un filtre passe-bas et utiliser les résultats obtenus au chapitre 5. En particulier,

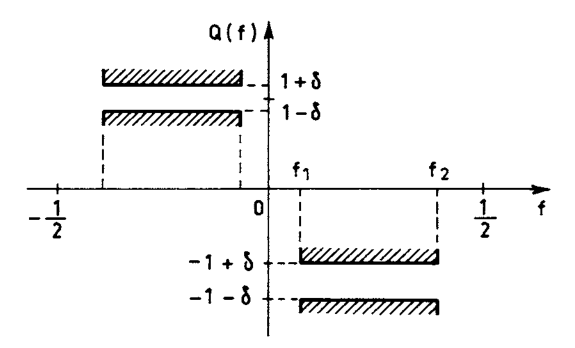


FIG. 9.6. *Gabarit de filtre de quadrature*

on peut faire appel au filtre demi-bande introduit au paragraphe 5.8, dont la réponse en fréquence est représentée sur la figure 9.7.

FIG. 9.7. *Réponse en fréquence d'un filtre demi-bande*

Ce filtre est spécifié par sa bande de transition Δf et les ondulations en bandes passante et affaiblie, qui sont égales à δ_0. Comme indiqué au paragraphe 5.8, les coefficients pairs sont nuls et la réponse en fréquence a pour expression, avec N = 4M + 1 coefficients :

$$H(f) = e^{-j2\pi 2Mf} \cdot \frac{1}{2}\left[1 + 2\sum_{i=1}^{M} h_{2i-1}\cos[2\pi(2i-1)f]\right] \qquad (9.28)$$

Une translation de cette réponse égale à 0,25 sur l'axe des fréquences conduit à la fonction $H'(f)$ telle que :

$$H'(f) = H(f-0{,}25) = e^{-j2\pi 2Mf}\,\frac{1}{2}\left[1 - 2\sum_{i=1}^{M}(-1)^i h_{2i-1}\sin[2\pi(2i-1)f]\right]$$

Les coefficients h'_n du filtre correspondant s'écrivent :

$$h'_{2i-1} = j.(-1)^i.h_{2i-1}; \quad h'_{-(2i-1)} = -j.(-1)^i.h_{2i-1}; \quad 1 \leqslant i \leqslant M$$

Ils prennent des valeurs imaginaires. En rapprochant l'expression $H'(f)$ de la relation (9.27) donnant $Q(f)$, il apparaît que l'ensemble des coefficients a_n tels que :

$$a_{-(2i-1)} = -a_{2i-1} = 2.(-1)^i.h_{2i-1}; \quad 1 \leqslant i \leqslant M$$

constitue l'ensemble des coefficients d'un filtre de quadrature dont l'ondulation est égale à $2\delta_0$ dans la bande $\left[\frac{\Delta f}{2}, \frac{1}{2} - \frac{\Delta f}{2}\right]$.

Exemple :

Pour les spécifications $\delta_0 = 0{,}01$ et $\Delta f = 0{,}111$ on trouve : M = 5.

$$a_1 = 0{,}6283$$
$$a_3 = 0{,}1880$$
$$a_5 = 0{,}0904$$
$$a_7 = 0{,}0443$$
$$a_9 = 0{,}0231$$

L'expression $H'(f)$ correspond à un filtre complexe qui comprend deux parties, d'une part un circuit de retard de 2M périodes élémentaires, d'autre part un circuit de filtre de quadrature, comme le montre la figure 9.8. Les sorties de ces deux circuits constituent la partie réelle et la partie imaginaire du signal complexe. On peut dire que le système comporte deux branches, une branche réelle et une branche imaginaire. Il permet finalement de convertir un signal réel en un signal analytique, c'est un filtre analytique RIF. Ce dispositif est parfois appelé modulateur IQ.

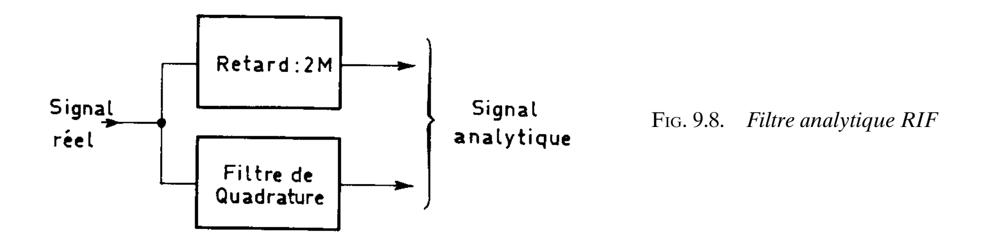


FIG. 9.8. *Filtre analytique RIF*

Il faut remarquer que la structure demeure si le filtre passe-bas de base n'est pas de type demi-bande, auquel cas les coefficients d'indice pair ne s'annulent pas. En fait, la translation de fréquence de 0,25 correspond à une multiplication des coefficients par un facteur complexe, tel que les coefficients h'_n prennent les valeurs :

$$h'_n = e^{-j\frac{\pi}{2}n} . h_n \tag{9.29}$$

Dans ces conditions la branche réelle du système n'est plus un simple retard ; une fonction de filtrage est réalisée en même temps que la génération du signal analytique.

Les circuits à réponse impulsionnelle finie permettent ainsi d'approcher le filtre de quadrature idéal sans faire d'erreur sur le déphasage mais en faisant une approximation de l'amplitude dans la bande passante. Les circuits à réponse impulsionnelle infinie ou récursifs fournissent une approche duale ; ils permettent, par l'utilisation de déphaseurs purs, d'approcher le filtre de quadrature sans erreur sur l'amplitude mais avec une approximation sur la phase.

9.4 DÉPHASEURS À 90° DE TYPE RÉCURSIF

Un circuit déphaseur récursif est caractérisé par le fait que le numérateur et le dénominateur de sa fonction de transfert en Z sont des polynômes images, c'est-à-dire qu'ils présentent les mêmes coefficients mais dans l'ordre inverse. Les propriétés des déphaseurs ont été introduites au paragraphe 6.3.

Il est possible de concevoir un couple de déphaseurs de telle sorte que les signaux de sortie présentent une différence de phase approchant 90° avec une erreur inférieure à ε, dans une bande de fréquence (f_1, f_2) donnée. Les techniques de calcul sont les mêmes que pour les filtres RII. La procédure pour aboutir à une différence de phase ayant un comportement de type elliptique est la suivante [5] :

– Détermination de l'ordre du circuit :

$$N = \frac{K(k_1)K(\sqrt{1-k^2})}{K(k)K(\sqrt{1-k_1^2})}$$

avec les valeurs de paramètres :

$$k = \frac{\operatorname{tg}(\pi f_1)}{\operatorname{tg}(\pi f_2)} \; ; \quad k_1 = \left[\frac{1 - \operatorname{tg}(\varepsilon/2)}{1 + \operatorname{tg}(\varepsilon/2)} \right]^2$$

– Détermination des zéros z_i de la fonction de transfert en Z :

$$A = Sn\left[\frac{(4i+1)\,K\left(\sqrt{1-k^2}\right)}{2N}, \sqrt{1-k^2} \right]$$

(Sn : fonction elliptique).

$$p_i = -\operatorname{tg}(\pi f_1)\,\frac{A}{\sqrt{1-A^2}}$$

$$z_i = \frac{1+p_i}{1-p_i} \quad \text{pour} \quad 0 \leqslant i \leqslant N-1$$

Exemple :

Soit les spécifications : $f_1 = 0{,}028$; $f_2 = 0{,}33$ et $\varepsilon = 1°$ il vient :

$$\operatorname{tg}(\pi f_1) = 0{,}0875 \; ; \quad k = 0{,}0505 \; ; \quad k_1 = 0{,}9657 \; ; \quad N \simeq 4{,}8.$$

En prenant N = 5 on obtient :

$$\begin{array}{ll} p_0 = -0{,}0395 & z_0 = 0{,}9240 \\ p_1 = -0{,}3893 & z_1 = 0{,}4396 \\ p_2 = -3{,}8360 & z_2 = -0{,}5864 \\ p_3 = -1{,}0039 & z_3 = -0{,}00197 \\ p_4 = -0{,}1509 & z_4 = 0{,}7377 \end{array}$$

Pour constituer le circuit les trois premiers zéros z_i sont affectés à une branche, les deux derniers à l'autre branche et la différence de phase en fonction de la fréquence est donnée par la fonction $\varphi(f)$ représentée sur la figure 9.9.

Les déphaseurs de types récursifs permettent d'obtenir deux signaux en quadrature. Il faut noter que dans cette opération, ils introduisent aussi une distorsion de phase qui est la même pour les deux signaux.

Ces circuits peuvent être utilisés dans les équipements de modulation et de multiplexage.

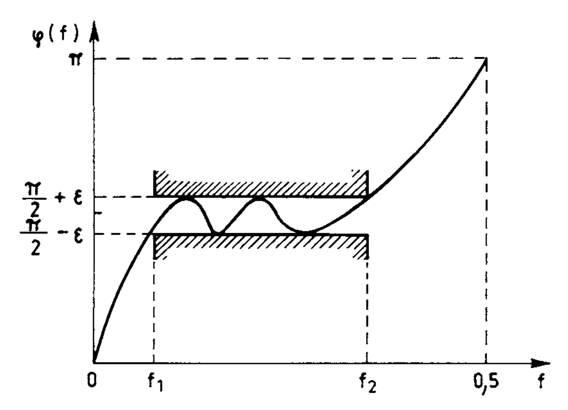


FIG. 9.9. *Caractéristique de déphaseurs à 90°*

9.5 MODULATION À BANDE LATÉRALE UNIQUE

La modulation d'un signal se traduit par un déplacement du spectre sur l'axe des fréquences. Elle est à bande latérale unique (BLU) si, pour un signal réel, la partie du spectre qui correspond aux fréquences positives est déplacée dans le sens des fréquences positives et la partie qui correspond aux fréquences négatives est déplacée vers les fréquences négatives.

Ainsi au signal : $s(t) = \cos \omega t$, correspond le signal modulé :

$$s_m(t) = \cos(\omega + \omega_0)t = \cos \omega t \cos \omega_0 t - \sin \omega t \sin \omega_0 t$$

Une telle opération peut être réalisée par la procédure suivante :

– Former le signal analytique $s_a(n) = s_R(n) + js_I(n)$ correspondant au signal réel que constitue la suite $s(n)$.

– Multiplier la suite $s_a(n)$ par la suite de nombres complexes :

$$\cos(2\pi n f_0) + j \sin(2\pi n f_0)$$

et conserver seulement la partie réelle $s_m(n)$ de la suite ainsi obtenue ; il vient :

$$s_m(n) = s_R(n) \cos(2\pi n f_0) - s_I(n) \sin(2\pi n f_0)$$

L'évolution du spectre du signal est donnée par la figure 9.10 et les circuits correspondants par la figure 9.11.

Si le filtre analytique est du type RIF, la suite $s_R(n)$ est simplement la suite $s(n)$ retardée. Avec les déphaseurs à 90° récursifs, il n'en est plus de même.

La suite correspondant au signal modulé $s_m(n)$ peut être additionnée à d'autres suites modulées pour fournir un signal multiplexé en fréquence comme en téléphonie par exemple.

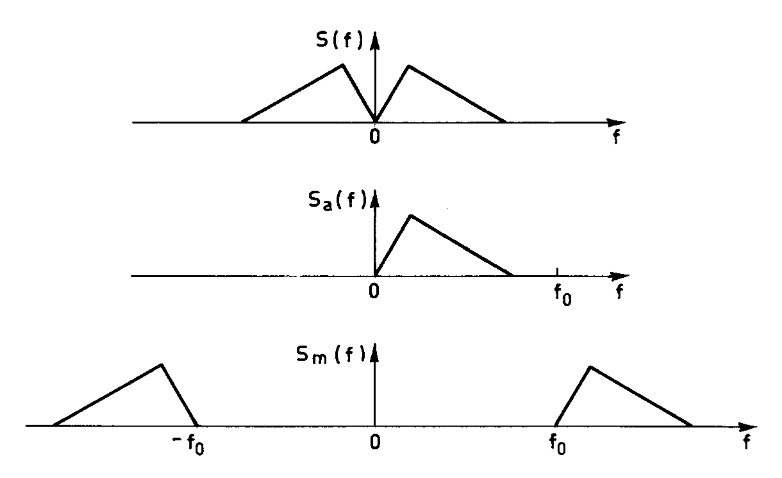


FIG. 9.10. *Modulation à bande latérale unique*

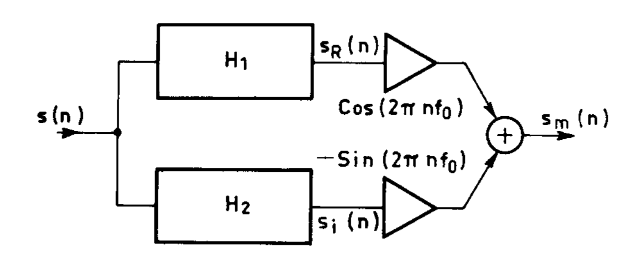


FIG. 9.11. *Circuit de modulation BLU*

9.6 LES FILTRES À DÉPHASAGE MINIMAL

Les propriétés des signaux causaux et analytiques étudiées au début de ce chapitre permettent d'éclaircir un point du calcul des filtres qui n'a pas été traité, concernant les caractéristiques de phase [3].

La réponse en fréquence d'un filtre $H(f)$ s'écrit :

$$H(f) = A(f)e^{-j\varphi(f)} \quad \text{avec} \quad A(f) = |H(f)|$$

et

$$\varphi(f) = -\text{Arg}[H(f)]$$

Le terme $A(f)$ est l'affaiblissement et $\varphi(f)$ est le déphasage apporté à un signal sinusoïdal, de fréquence f, par le filtre.

Si le filtre est à coefficients réels, sa réponse impulsionnelle, la suite $h(n)$, est réelle et par suite :

$$H(f) = \overline{H(-f)}; \quad A(f) = A(-f) \quad \text{et} \quad \varphi(f) = -\varphi(-f)$$

Le terme $h(n)$ est obtenu par :

$$h(n) = 2 \int_0^{\frac{1}{2}} \mathrm{A}(f) \cos [2\pi n f - \varphi(f)] \, df \qquad (9.30)$$

La réponse ne pouvant précéder l'application du signal au filtre, un filtre réalisable est nécessairement causal, avec :

$$h(n) = 0 \quad \text{pour} \quad n < 0$$

Il s'en suit que si la réponse est décomposée en partie réelle et partie imaginaire :

$$\mathrm{H}(f) = \mathrm{H_R}(f) + j\mathrm{H_I}(f)$$

les fonctions $\mathrm{H_R}(f)$ et $\mathrm{H_I}(f)$ sont liées par les relations (9.11) et (9.12) données au paragraphe 9.1.

Or un filtre est souvent spécifié seulement par la donnée de contraintes sur l'amplitude :

$$\mathrm{A}^2(f) = \mathrm{H_R^2}(f) + \mathrm{H_I^2}(f)$$

et il en résulte une indétermination sur le déphasage.

D'une manière générale, le traitement d'un signal demande un certain temps, qui correspond au temps de propagation à travers le système. Ce paramètre est caractérisé par le déphasage en fonction de la fréquence. Pour minimiser ce temps de propagation, on est conduit à rechercher la caractéristique de déphasage minimal, ce qui lève l'indétermination sur le calcul du filtre. Une autre possibilité pour lever cette indétermination est de spécifier un déphasage linéaire.

Un filtre stable et réalisable a une fonction de transfert en Z, H(Z), dont les pôles sont à l'intérieur du cercle unité, les zéros pouvant être à l'extérieur. Soit $\mathrm{Z_0}$ un tel zéro et soit $\mathrm{H_1(Z)}$ la fonction telle que :

$$\mathrm{H_1(Z)} = \frac{1 - 2\,\mathrm{Re}(\mathrm{Z_0})\mathrm{Z}^{-1} + |\mathrm{Z_0}|^2\mathrm{Z}^{-2}}{|\mathrm{Z_0}|^2 - 2\,\mathrm{Re}(\mathrm{Z_0})\mathrm{Z}^{-1} + \mathrm{Z}^{-2}} = \mathrm{Z}^{-2}\,\frac{(\mathrm{Z} - \mathrm{Z_0})(\mathrm{Z} - \overline{\mathrm{Z}}_0)}{(\mathrm{Z}^{-1} - \overline{\mathrm{Z}}_0)(\mathrm{Z}^{-1} - \mathrm{Z_0})}$$

C'est la fonction de transfert d'un déphaseur pur du second ordre, qui a été introduite au paragraphe 6.3 et dont le temps de propagation de groupe est donné par la relation (6.46). On peut écrire :

$$\mathrm{H(Z)} = \mathrm{H_1(Z)} . \mathrm{H_2(Z)}$$

où $\mathrm{H_2(Z)}$ est une fonction qui s'annule en Z_0^{-1} et apporte un déphasage plus faible que H(Z). Un raisonnement par itération amène à la conclusion que la fonction qui a le déphasage minimal est la fonction $\mathrm{H}_m\mathrm{(Z)}$ obtenue en remplaçant dans H(Z) tous les zéros extérieurs au cercle unité par leur inverse.

La formulation de la condition de phase minimale est que la fonction :

$$\ln[\mathrm{H(Z)}] = \ln[\mathrm{A(Z)}] - j\varphi(\mathrm{Z})$$

n'ait pas de pôles à l'extérieur du cercle unité.

Dans ces conditions les fonctions ln [A(*f*)] et φ(*f*) sont liées par les relations (9.11) et (9.12) correspondant à la transformation de Hilbert. Ce sont les relations de Bayard-Bode pour les systèmes discrets :

$$\ln [A(f)] = K - \int_{-\frac{1}{2}}^{\frac{1}{2}} \varphi(f') \operatorname{cotg} \pi(f-f')\, df' \qquad (9.31)$$

$$\varphi(f) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln [A(f')] \operatorname{cotg} \pi(f-f')\, df' \qquad (9.32)$$

La constante K est un facteur d'échelle pour l'amplitude.

Une autre formulation de la condition de déphasage minimal, pour le filtre réalisable et stable défini par H(Z), est que $H^{-1}(Z)$ corresponde également à un filtre réalisable et stable. Un exemple est donné au paragraphe 15.3 avec la prédiction linéaire.

De même, une fonction $H_M(Z)$ dont les zéros sont à l'extérieur du cercle unité est dite à déphasage maximal.

9.7 FILTRE DIFFÉRENTIATEUR

Un filtre différentiateur est un filtre de quadrature dont la réponse est proportionnelle à la fréquence :

$$H_d(\omega) = D(\omega)\, e^{-j\frac{\pi}{2}} \qquad (9.33)$$

$$D(\omega) = \omega \quad \text{pour} \quad 0 \leqslant \omega \leqslant \omega_1 \leqslant \pi.$$

Si la fin de bande passante ω_1 est égale à π, le filtre est dit pleine bande.

Le filtre numérique à N coefficients correspondant a pour réponse :

$$H(\omega) = R(\omega) e^{-j\left(\frac{\pi}{2} + \omega(N-1)/2\right)} \qquad (9.34)$$

avec R(ω) fonction réelle telle que

$$R(\omega) = \sum_{i=1}^{P} h_i \sin i\omega; \quad N = 2P + 1 \qquad (9.35)$$

$$R(\omega) = \sum_{i=1}^{P} h_i \sin\left(i - \frac{1}{2}\right)\omega; \quad N = 2P.$$

Les réponses impulsionnelles de ce type de filtre ont été présentées au paragraphe 5.2.

Les méthodes de calcul des coefficients données pour les filtres généraux s'appliquent à ce cas particulier, notamment les moindres carrés. Un cas simple est celui du filtre pleine bande, qui est nécessairement à nombre de coefficients pair

comme le montre la relation (9.35) et pour lequel la technique des moindres carrés conduit à l'expression suivante pour les coefficients [6] :

$$h_i = \frac{8}{\pi} \frac{(-1)^{i+1}}{(2i-1)^2} \tag{9.36}$$

Si la fonction désirée D(ω) est proportionnelle à une puissance de la fréquence, le différentiateur est dit d'ordre supérieur.

La conversion d'un signal réel en signal complexe s'accompagne souvent d'une opération d'interpolation, notamment dans les interfaces analogique-numériques des récepteurs de communication.

9.8 INTERPOLATION PAR FILTRE RIF

L'interpolation consiste à calculer certaines valeurs du signal entre les échantillons connus. C'est une fonction de filtrage qu'il est généralement commode de réaliser par des filtres RIF [7, 8].

Soit à calculer les valeurs $x(nT + \tau)$, à partir de la suite $x(nT)$ en utilisant un filtre RIF à N = 2P + 1 coefficients. Le retard τ est tel que $|\tau| \leqslant T/2$.

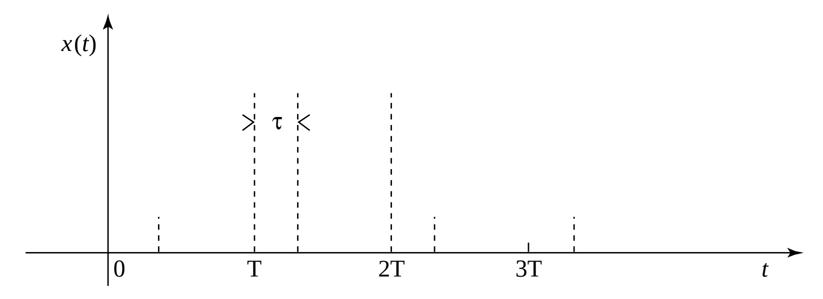


FIG. 9.12. *Interpolation avec un retard τ*

Le filtre lui-même apporte un retard KT, avec K entier et la sortie $y(n)$ doit être telle que :

$$y(n) \approx x[(n-K)T + \tau] \tag{9.37}$$

La fonction de retard s'exprime également avec la fonction de transfert en Z. Il faut avoir, en posant T = 1 :

$$H(Z) = \sum_{i=0}^{N-1} a_i Z^{-i} \approx Z^{-K+\tau} \tag{9.38}$$

ou encore dans le domaine des fréquences :

$$\left(\sum_{i=0}^{N-1} a_i e^{-ji\omega}\right) e^{j(K-\tau)\omega} \approx 1 \tag{9.39}$$

soit :

$$e^{-j\tau\omega} \sum_{i=0}^{N-1} a_i \, e^{-j(i-K)\omega} \approx 1 \tag{9.40}$$

En posant K = P, après changement de variable, il vient :

$$e^{-j\tau\omega} \sum_{i=-P}^{P} h_i \, e^{-ji\omega} \approx 1 \tag{9.41}$$

et, finalement :

$$G(\omega) = \sum_{i=-P}^{P} h_i \, e^{-j(i+\tau)\omega} \approx 1 ; \quad |\tau| \leqslant \frac{1}{2} \tag{9.42}$$

Les coefficients h_i de l'interpolateur se déterminent à partir de cette relation et on peut utiliser les techniques d'approximation classiques, par exemple les moindres carrés.

Pour les systèmes dans lesquels le retard peut varier, comme les boucles de synchronisation, il est intéressant de pouvoir relier les valeurs des coefficients aux valeurs du retard τ, ce qui permet à l'interpolateur de suivre l'évolution du retard. L'interpolation de Lagrange est une approche de ce type.

9.9 INTERPOLATION DE LAGRANGE

Dans le domaine fréquentiel, l'interpolation de Lagrange correspond à un filtrage «max flat», c'est-à-dire avec annulation des dérivées de la réponse en fréquence à l'origine. Les coefficients sont obtenus par résolution du système d'équations linéaires suivant :

$$G(0) = 1$$

$$G^{(p)}(0) = 0 ; \quad 1 \leqslant p \leqslant P \tag{9.43}$$

Ainsi pour P = 1, on obtient :

$$h_{-1} + h_0 + h_1 = 1$$

$$(\tau - 1)h_{-1} + \tau h_0 + (\tau + 1)h_1 = 0 \tag{9.44}$$

$$(\tau - 1)^2 h_{-1} + \tau^2 h_0 + (\tau + 1)^2 h'_1 = 0$$

ce qui conduit à l'équation matricielle :

$$\begin{bmatrix} 1 & 1 & 1 \\ \tau - 1 & \tau & \tau + 1 \\ (\tau - 1)^2 & \tau^2 & (\tau + 1)^2 \end{bmatrix} \begin{bmatrix} h_{-1} \\ h_0 \\ h_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \tag{9.45}$$

dont la solution s'écrit :

$$h_{-1} = \frac{\tau(\tau+1)}{2}; \quad h_0 = 1-\tau^2; \quad h_1 = \frac{(\tau-1)\tau}{2}$$

On peut observer que la norme de la fonction H(ω) s'écrit :

$$\|H\|_2^2 = \sum_{i=-P}^{P} h_i^2 = 1 - \frac{3}{2}\tau^2 + \frac{3}{2}\tau^4 \tag{9.46}$$

ce qui montre que l'erreur quadratique d'interpolation croît avec le retard τ et est maximale pour $\tau = \frac{1}{2}$.

Une mise en œuvre efficace est obtenue en remarquant que le système (9.45) conduit à des coefficients qui peuvent se mettre sous la forme :

$$h_i = \sum_{j=0}^{N-1} b_{ij}\tau^j \tag{9.47}$$

avec $b_{00} = 1$; $b_{i0} = 0$ pour $i \neq 0$. Il vient alors :

$$Z^P H(Z) = \sum_{i=-P}^{P} \left(\sum_{j=0}^{N-1} b_{ij}\tau^j \right) Z^{-i}$$

et, en inversant les sommations :

$$Z^P\, H(Z) = \sum_{j=0}^{N-1} C_j(Z)\, \tau^j \tag{9.48}$$

avec :

$$C_0 = 1; \quad C_j(Z) = \sum_{i=-P}^{P} b_{ij} Z^{-i}$$

Le schéma de réalisation correspondant est donné par la figure 9.13. Il s'adapte facilement aux évolutions du retard τ.

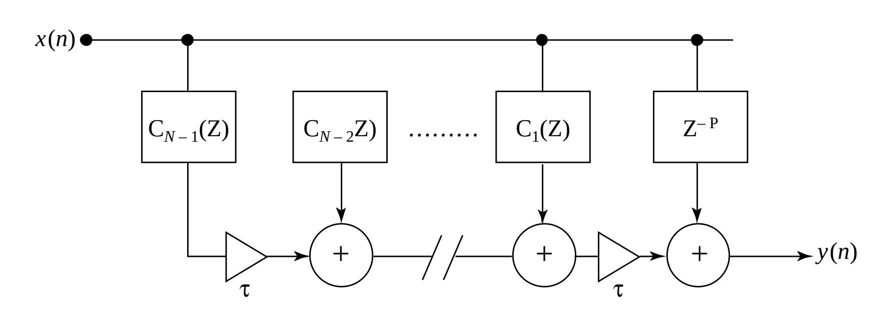


Fig. 9.13. Réalisation d'un interpolateur de Lagrange

L'expression générale des coefficients h_i, solution du système (9.45) est obtenue en remarquant que la matrice carrée est de type Vandermonde. Les détermi-

nants et sous-déterminants s'annulent quand deux lignes ou deux colonnes sont identiques, il en résultent qu'ils se mettent sous la forme de produits de facteurs.

En généralisant à des retards Δ quelconques, les coefficients du filtre d'interpolation de Lagrange s'écrivent :

$$a_i = \prod_{j=0;\, j \neq 1}^{N-1} \frac{\Delta - j}{i - j}; \quad 0 \leqslant i \leqslant N-1 \tag{9.49}$$

et les valeurs interpolées :

$$x(nT - \Delta) \approx y(n) = \sum_{i=0}^{N-1} a_i\, x[(n-i)T] \tag{9.50}$$

La qualité de l'interpolation est fonction du nombre coefficients. Quand ce nombre tend vers l'infini, on retrouve la formule de l'échantillonnage (1.57), en utilisant l'identité :

$$\sin \pi t = \pi t \prod_{n \neq 0} \left(1 - \frac{t}{n}\right) \tag{9.51}$$

qui permet d'écrire :

$$a_i = \prod_{j \neq i} \frac{\Delta - j}{i - j} = \prod_{j \neq 0} \frac{k - (\Delta - i)}{k} = \frac{\sin \pi\, (\Delta - i)}{\pi(\Delta - i)} \tag{9.52}$$

et en posant $\Delta = \dfrac{t}{T}$.

Dans certaines applications, comme le tracé de courbes ou le traitement d'images, les données sont disponibles par bloc et il faut interpoler des valeurs dans le bloc.

9.10 INTERPOLATION PAR BLOC – SPLINES

Le fait d'avoir à traiter un ensemble fini de données permet de faire appel à des filtres dont la réponse impulsionnelle ne possède pas la propriété de Nyquist, c'est-à-dire qu'elle ne s'annule pas aux multiples entiers de la période d'échantillonnage. Mais alors, pour conserver les échantillons connus, il faut effectuer un prétraitement des données, qui est une opération de filtrage inverse. Les fonctions de ce type les plus utilisées sont les splines [9].

La fonction spline de degré m est définie par les m convolutions suivantes :

$$B_m(t) = B_0(t) * B_0(t) * \ldots * B_0(t) \tag{9.53}$$

où $B_0(t)$ est l'impulsion de largeur unité :

$$B_0(t) = \begin{cases} 1 & ; \quad |t| < \dfrac{1}{2} \\ 0{,}5 & ; \quad |t| = \dfrac{1}{2} \\ 0 & ; \quad |t| > \dfrac{1}{2} \end{cases}$$

La convolution étant une intégration, plus le degré est élevé plus la fonction est lisse, puisqu'il faut aller de plus en plus loin dans l'ordre des dérivées pour trouver une discontinuité. Une fonction très utilisée est la spline cubique.

$$B_3(t) = \begin{cases} \dfrac{2}{3} - |t|^2 + \dfrac{|t|^3}{2} & ; \quad 0 \leqslant |t| < 1 \\ \dfrac{2 - |t|^3}{6} & ; \quad 1 \leqslant t < 2 \\ 0 & ; \quad 2 \leqslant |t| \end{cases} \tag{9.54}$$

Étant donné un ensemble de N valeurs $s(nT)$, avec $0 \leqslant n \leqslant N - 1$, il faut d'abord effectuer une opération de filtrage inverse et obtenir un nouvel ensemble de valeur $x(n)$, auquel sera appliqué le filtre d'interpolation.

Pour la spline cubique, avec $T = 1$, les échantillons ont pour transformée en Z :

$$B_3(Z) = \frac{Z + 4 + Z^{-1}}{6} \tag{9.55}$$

et il vient :

$$B_3^{-1}(Z) = \frac{6}{2 + \sqrt{3}} \, \frac{1}{1 + (2 - \sqrt{3})Z^{-1}} \cdot \frac{1}{1 + (2 - \sqrt{3})Z} \tag{9.56}$$

ou encore :

$$B_3^{-1}(Z) = \frac{6 - 3\sqrt{3}}{2\sqrt{3} - 3} \left[\frac{1}{1 + (2 - \sqrt{3})Z^{-1}} + \frac{1}{1 + (2 - \sqrt{3})Z} - 1 \right] \tag{9.57}$$

Des deux facteurs de $B_3^{-1}(Z)$ dans (9.56), l'un est causal et l'autre anti-causal. Le premier facteur peut être appliqué à la suite $s(n)$ pour fournir la sortie $u(n)$.

$$u(n) = s(n) - (2 - \sqrt{3})u(n - 1) \tag{9.58}$$

Le second facteur, appliqué à la suite $u(n)$ donne la suite $x(n)$ cherchée :

$$x(n) = \frac{6}{2 + \sqrt{3}} \cdot u(n) - (2 - \sqrt{3})x(n + 1) \tag{9.59}$$

Dans ce cas, il faut effectuer les calcul dans la direction inverse.

Les deux récurrences demandent des valeurs initiales, $u(0)$ et $x(\mathrm{N}-1)$, dont la détermination dépend des conditions aux limites du bloc de données. Généralement, on considère une extension par symétrie en dehors du bloc de données, c'est-à-dire $s(-n) = s(n)$ et $s(\mathrm{N}-1+k) = s(\mathrm{N}-1-k)$. La périodicité est alors de 2N – 2 et le développement en série du premier facteur dans l'expression (9.56) conduit à la valeur initiale :

$$u(0) = \sum_{n=0}^{\infty} (\sqrt{3}-2)^n s(n)$$

soit :

$$u_0 = \frac{1}{1-(\sqrt{3}-2)^{2\mathrm{N}-2}} \sum_{n=0}^{2\mathrm{N}-3} (\sqrt{3}-2)^n s(n) \tag{9.60}$$

Pour les grandes valeurs de N et selon la précision recherchée, la sommation peut être limitée aux premiers termes.

Dans l'autre sens, la valeur de $x(\mathrm{N}-1)$ peut être obtenue par calcul direct. En effet, par la décomposition de $\mathrm{B}_3^{-1}(\mathrm{Z})$ en fractions rationnelles donnée par (9.57), puis un développement en série des deux termes et en utilisant l'expression de $u(n)$ en fonction de $s(n)$ correspondante pour $n = \mathrm{N}-1$, avec la symétrie de $s(n)$ autour de N – 1, on aboutit à l'initialisation suivante :

$$x(\mathrm{N}-1) = \frac{6-3\sqrt{3}}{2\sqrt{3}-3}[2u(\mathrm{N}-1) - s(\mathrm{N}-1)] \tag{9.61}$$

On vérifie bien que pour le signal constant $s(n) = 1$, il vient :

$$u(0) = \frac{1}{3-\sqrt{3}} = u(n) \text{ et } x(\mathrm{N}-1) = 1 = x(n).$$

Une fois obtenues les valeurs du bloc transformé, l'interpolation est effectuée par

$$s(t) = \sum_n x(n)\,\mathrm{B}_m(t-n\mathrm{T}) \tag{9.62}$$

La sommation est limitée aux termes pour lesquels la fonction spline n'est pas nulle. Quand le degré n croît, comme au paragraphe précédent, cet interpolateur tend vers l'interpolateur idéal.

9.11 CONCLUSION

La transformation de signaux réels en signaux complexes est une opération de filtrage dit de quadrature. Ce filtre de quadrature peut avoir une réponse en fréquence quelconque et il permet, notamment, de réaliser une réponse proportionnelle à la fréquence, c'est le filtre différentiateur.

En pratique, les transformations réel-complexe et inversement se réalisent efficacement par interpolation à l'aide d'un filtre demi-bande. Le gabarit de ce filtre se définit à partir des objectifs de performance sur la distorsion de fréquence et les résidus de bande image.

L'interpolation est une opération fondamentale liée à l'échantillonnage.

En théorie, elle se définit par la formule de l'échantillonnage qui correspond à un filtre à phase linéaire et à réponse impulsionnelle infinie.

En pratique, pour conserver les échantillons connus, il faut faire appel à un filtre RIF à phase linéaire. Un cas particulier important est l'interpolation de Lagrange qui correspond à un filtre dit « max flat », dont les dérivées de la réponse en fréquence s'annulent à l'origine.

L'interpolation par bloc, comme en traitement d'images par exemple, peut utiliser des réponses de filtres qui ne conservent pas les échantillons connus, comme les fonctions splines, moyennant un prétraitement.

Les fonctions splines constituent une autre approximation de l'interpolateur idéal.

BIBLIOGRAPHIE

[1] E. Roubine – *Introduction à la théorie de la communication.* Tome I, Ed. Masson, 1970.

[2] B. Picinbono – *Éléments de théorie du signal.* Dunod Université, 1977.

[3] A. Oppenheim and S. Schafer – *Digital Signal Processing.* Chapter 7, Prentice Hall, N.J., 1974.

[4] S. Mitra, J. F. Kaiser – *Handbook for Digital Signal Processing*, John Wiley, New-York, 1993.

[5] B. Gold and C. Rader – *Digital Processing of Signals.* Chapter 3, Mc Graw-Hill, 1969.

[6] G. Mollova – *Compact Formulas for Least Squares Design of Digital Differentiators*, Electronics Letters, Vol. 35, N° 20, 1999, pp. 1695-97.

[7] T. I. Laakso, V. Valimaki, M. Karjalainen, U. K. Laine, *Splitting the unit delay – Tools for fractional delay filter design*, IEEE Signal Processing Magazine, Vol. 13, N° 1, pp. 30-60, Janv. 1996.

[8] J. J. Fuchs, B. Delyon, *Minimum L1-norm reconstruction Function for oversampled signals : application to time delay estimation*, IEEE Trans. Vol. IT-46, July 2000, pp. 1666-73.

[9] M. Unser, *Splines : a Perfect Fit for Signal and Image Processing*, IEEE Signal Processing Magazine, Vol. 16, N° 6, November 1999, pp. 22-38.

EXERCICES

1 Calculer la Transformée de Fourier $X(f)$ de la suite réelle et causale $x(n)$ telle que :

$$x(n) = 0 \quad \text{pour} \quad n < 0$$
$$x(n) = a^n \quad \text{pour} \quad n \geqslant 0$$

avec

$$|a| < 1$$

Décomposer $X(f)$ en parties réelle et imaginaire.

2 Montrer que la fonction $X(Z)$ telle que :

$$X(Z) = \frac{1}{1 - aZ^{-1}}$$

peut être obtenue à partir de sa partie réelle sur le cercle unité, la fonction $X_R(\omega)$ donnée par :

$$X_R(\omega) = \frac{1 - a\cos\omega}{1 - 2a\cos\omega + a^2} \quad \text{avec} \quad |a| < 1$$

3 A partir d'un signal réel représenté par la suite $x(n)$, on forme un signal complexe dont les parties réelle et imaginaire sont données par :

$$x_R(n) = x(n)\cos\left(2\pi\frac{n}{4}\right)$$

$$x_I(n) = x(n)\sin\left(2\pi\frac{n}{4}\right)$$

Quelles remarques peut-on faire sur les suites $x_R(n)$ et $x_I(n)$?

Le signal obtenu est-il un signal analytique ?

Un filtre demi-bande est appliqué à chacune des suites $x_R(n)$ et $x_I(n)$ et le signal complexe ainsi filtré est multiplié par la suite complexe $e^{-j2\pi\frac{n}{4}}$. Quelle opération a été ainsi effectuée sur le signal réel $x(n)$? Effectuer la suite de ces opérations sur le signal $x(n) = \cos\left(\pi\frac{n}{5}\right)$.

4 Étudier l'incidence sur un filtre de quadrature RIF de la limitation du nombre de bits des coefficients. En reprenant la démarche des paragraphes 5.6 et 5.8 pour les filtres RIF à phase linéaire, rechercher une formule d'estimation donnant le nombre de bits des coefficients en fonction des paramètres du filtre de quadrature, ondulation d'amplitude et bande de transition.

5 Établir une expression simplifiée pour l'ordre d'un déphaseur à 90° de type RII. Étudier l'incidence de la limitation du nombre de bits des coefficients sur les caractéristiques et rechercher une formule d'estimation en fonction des paramètres. Effectuer une vérification des résultats obtenus sur l'exemple du paragraphe 9.4.

6 Soit la fonction $H(Z)$ définie par :

$$H(Z) = [(1 - Z_0Z^{-1})(1 - \overline{Z}_0Z^{-1})]^2$$

avec

$$Z_0 = 0{,}5\,(1 + j)$$

Cette fonction est à phase minimale. Donner l'expression de la fonction à phase linéaire et de la fonction à phase maximale qui ont la même caractéristique d'amplitude. Comparer les réponses impulsionnelles.

7 Pour convertir un signal réel échantillonné à la fréquence $2f_e$ en un signal complexe échantillonné à la fréquence f_e, on utilise un filtre demi-bande de fonction de transfert

$$H(Z) = -\,0{,}0506 + 0{,}2954\ Z^{-2} + 0{,}5\ Z^{-3} + 0{,}2951\ Z^{-4} - 00506\ Z^{-6}$$

Calculer la réponse du filtre aux fréquences 0 et $f_e/8$. En déduire l'ondulation et la largeur de la bande de transition.

On utilise ce filtre pour réaliser un modulateur IQ auquel on applique le signal $x(n) = \sin\left(n\,\dfrac{\pi}{4}\right)$. Donner l'expression du signal complexe $y(n)$ en sortie et montrer qu'il comporte une composante à la fréquence $f_e/4$ et une composante à la fréquence $3f_e/4$. Quelles sont les amplitudes de ces composantes ?

8 Dans un récepteur à modulation de fréquence, le discriminateur (convertisseur amplitude-fréquence) est un filtre différenciateur à 5 coefficients ayant les valeurs suivantes :

$$h = [-\,0{,}1766 \quad 0{,}9696 \quad 0 \quad -\,0{,}9696 \quad 0{,}1766]$$

Tracer la réponse de ce filtre et déterminer la bande passante et l'ondulation en bande passante.

Quel est le retard apporté par ce filtre ?

Chapitre 10

Le filtrage multicadence

Le filtrage multicadence est une technique qui a pour objet de réduire la vitesse de calcul dans les filtres numériques et en particulier le nombre de multiplications à faire par seconde. En effet, ce paramètre est généralement considéré comme représentatif de la complexité des systèmes.

Dans un filtre, le nombre de multiplications à faire par seconde M_R s'exprime par :

$$M_R = K . f_e$$

où f_e est la cadence à laquelle se font les calculs. Le paramètre f_e correspond généralement à la fréquence d'échantillonnage du signal que représentent les nombres traités. Le facteur K dépend du type de filtre et de ses performances.

Pour réduire la valeur de M_R on peut agir sur le facteur K, en choisissant le type et la structure de filtre les mieux appropriés et en optimisant l'ordre de ce filtre en fonction des contraintes et des caractéristiques à obtenir. On peut également agir sur l'autre facteur, f_e, en faisant varier la fréquence d'échantillonnage au cours du traitement; les avantages ainsi obtenus sont considérables dans de nombreux cas pratiques.

La fréquence d'échantillonnage d'un signal réel est supérieure au double de sa largeur de bande. Au cours du traitement la largeur de bande varie; par exemple, une opération de filtrage élimine les composants indésirables et la largeur de bande utile se trouve réduite. En un point où la bande utile a été réduite, la fréquence d'échantillonnage du signal peut elle-même être réduite. Il s'en suit que la fréquence d'échantillonnage peut être adaptée à la largeur de bande du signal à chaque étape du traitement, pour minimiser la vitesse de calcul dans un filtre. Avant d'étudier les développements et la mise en œuvre de ce principe de base, il convient d'abord d'analyser l'incidence sur le signal et son spectre d'un changement de fréquence d'échantillonnage.

10.1 SOUS-ÉCHANTILLONNAGE ET TRANSFORMÉE EN Z

Comme l'augmentation de fréquence d'échantillonnage, ou interpolation, la réduction de fréquence d'échantillonnage, ou décimation, modifie la transformée en Z du signal. Le cas d'une réduction par le facteur 2 est illustré par la figure 10.1, qui montre les opérations effectuées sur le signal, le symbole de décimation et les spectres. La suppression d'un échantillon sur deux est obtenue en ajoutant la suite d'origine et la même suite après inversion du signe d'un échantillon sur deux. Cette inversion décale le spectre autour de la demi-fréquence d'échantillonnage. Après addition des spectres, la périodicité dans le domaine des fréquences

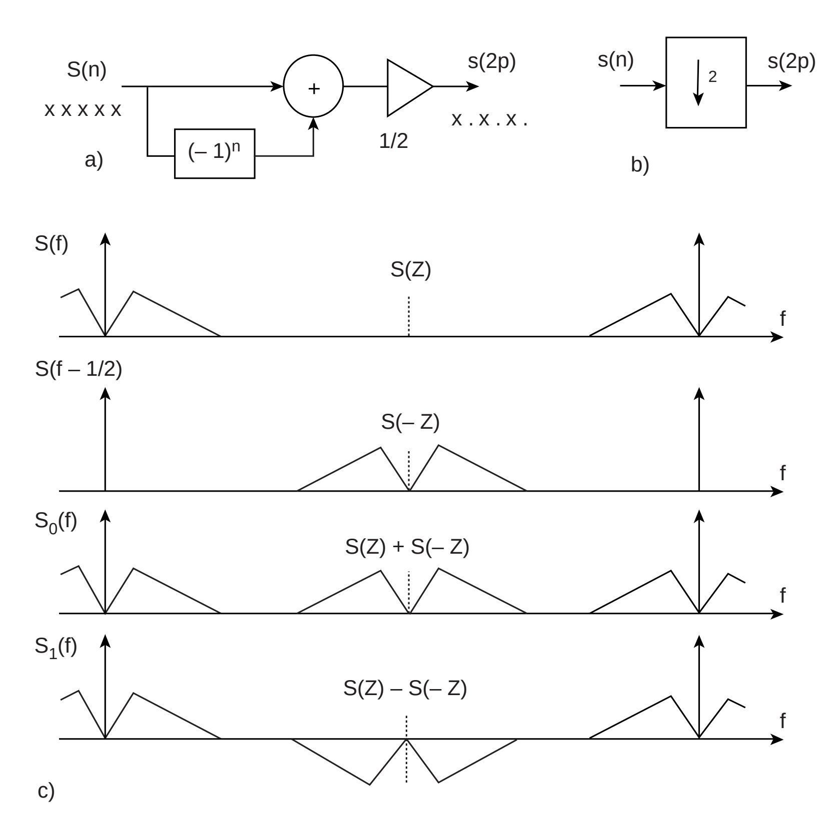


FIG. 10.1. *Sous-échantillonnage par 2*
a) opérations de décimation
b) symbole de la décimation
c) spectres et transformées en Z

se trouve divisée par deux. Quand on décompose la suite s(n) en deux suites entrelacées, en posant :

$$S_0(Z^2) = [S(Z) + S(-Z)]/2 \quad ; \quad Z^{-1}S_1(Z^2) = [S(Z) - S(-Z)]/2$$

on obtient les transformées en Z des deux suites, le terme Z^2 caractérisant le doublement de la période d'échantillonnage et le facteur Z^{-1} représentant l'entrelacement. La transformée en Z de s(n) est reconstituée par :

$$S(Z) = S_0(Z^2) + Z^{-1} S_1(Z^2)$$

Les formules de décomposition et reconstitution établies pour le facteur 2 se généralisent à un entier M quelconque. Cette généralisation est abordée par les transformées de Fourier.

Soit le signal $s(t)$ dont le spectre $S(f)$ ne contient pas de composantes aux fréquences supérieures à une valeur f_m et supposé échantillonné avec la période T telle que :

$$\frac{1}{MT} > 2f_m; \quad \text{M entier}$$

On se propose d'examiner la relation qui existe entre les transformées de Fourier $S_i(f)$ des suites entrelacées :

$$s\left[\left(n + \frac{i}{M}\right) MT\right]; \quad i = 0, 1, 2, \ldots, M-1.$$

D'après les résultats du paragraphe 1.2, la transformée de Fourier de la distribution $u_0(t)$, telle que :

$$u_0(t) = \sum_{n=-\infty}^{\infty} \delta(t - nMT)$$

est la distribution $U_0(f)$ telle que :

$$U_0(f) = \sum_{n=-\infty}^{\infty} e^{-j2\pi f nMT} = \frac{1}{MT} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{MT}\right)$$

La transformée de Fourier de la distribution $u_i(t)$, telle que :

$$u_i(t) = \sum_{n=-\infty}^{\infty} \delta\left[t - \left(n + \frac{i}{M}\right) MT\right]; \quad i = 0, 1, \ldots, M-1. \qquad (10.1)$$

est la distribution $U_i(f)$ telle que :

$$U_i(f) = \sum_{n=-\infty}^{\infty} e^{-j2\pi f\left(n + \frac{i}{M}\right)MT} = \frac{1}{MT} \cdot e^{-j2\pi f iT} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{MT}\right)$$

ou encore :

$$U_i(f) = \frac{1}{MT} \sum_{n=-\infty}^{\infty} e^{-j2\pi \frac{in}{M}} \delta\left(f - \frac{n}{MT}\right) \qquad (10.2)$$

Comme $S_i(f)$ ($i = 0, 1, \ldots, M-1$) est le produit de convolution de $S(f)$ par la distribution $U_i(f)$, il vient :

$$S_i(f) = \frac{1}{MT} \sum_{n=-\infty}^{\infty} e^{-j2\pi \frac{in}{M}} S\left(f - \frac{n}{MT}\right) \qquad (10.3)$$

Calculons maintenant le spectre $S^M(f)$ tel que :

$$S^M(f) = \sum_{i=0}^{M-1} S_i(f) = \frac{1}{MT} \sum_{n=-\infty}^{\infty} S\left(f - \frac{n}{MT}\right) \sum_{i=0}^{M-1} e^{-j2\pi \frac{in}{M}}$$

Comme la deuxième sommation s'annule pour toute valeur de n sauf les multiples de M, il vient :

$$S^M(f) = \frac{1}{T} \sum_{n=-\infty}^{\infty} S\left(f - \frac{n}{T}\right) \qquad (10.4)$$

spectre qui correspond à un signal échantillonné à la fréquence $\frac{1}{T}$.

Les termes $S_i(f)$ s'expriment également en fonction de $S^M(f)$. En effet dans la relation (10.3) la sommation peut se décomposer comme suit :

$$S_i(f) = \frac{1}{MT} \sum_{n=-\infty}^{\infty} \sum_{m=0}^{M-1} S\left[f - \left(n + \frac{m}{M}\right)\frac{1}{T}\right] e^{-j2\pi \frac{im}{M}}$$

ou encore :

$$S_i(f) = \frac{1}{M} \sum_{m=0}^{M-1} e^{-j2\pi \frac{im}{M}} \frac{1}{T} \sum_{n=-\infty}^{\infty} S\left[\left(f - \frac{m}{MT}\right) - \frac{n}{T}\right]$$

et finalement :

$$S_i(f) = \frac{1}{M} \sum_{m=0}^{M-1} e^{-j2\pi \frac{im}{M}} S^M\left(f - \frac{m}{MT}\right) \qquad (10.5)$$

La figure 10.2 illustre le cas où M = 4.

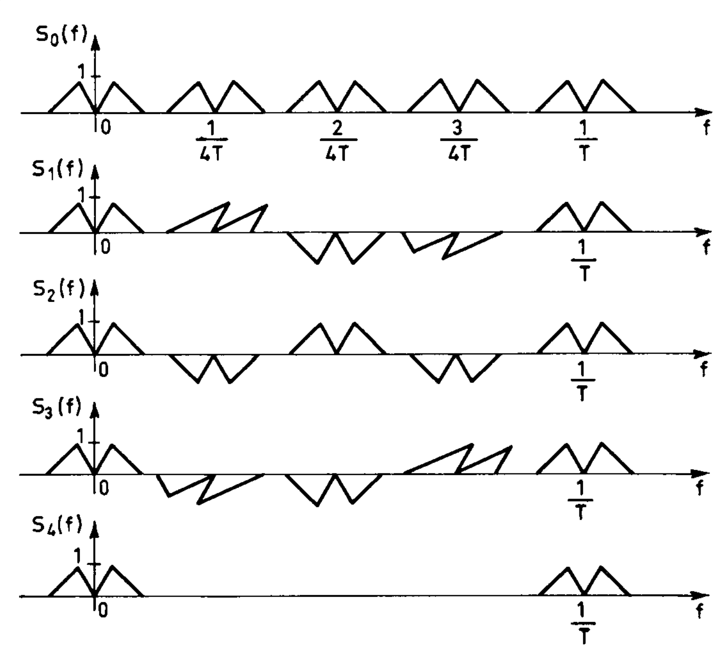


FIG. 10.2. *Spectres obtenus par 4 échantillonnages entrelacés*

Les spectres $S_i(f)$ correspondent à des échantillonnages entrelacés, à la fréquence $\frac{1}{MT}$ et le spectre $S^M(f)$ correspond à l'échantillonnage à la fréquence $\frac{1}{T}$. Le changement de fréquence d'échantillonnage revient à échanger ces spectres.

Il est intéressant de remarquer sur la figure 10.2 que le fait de retarder la suite des impulsions d'échantillonnage provoque des rotations de phase de valeurs multiples de $\frac{2\pi}{M}$ des bandes image autour des multiples de la fréquence d'échantillonnage $\frac{1}{MT}$.

L'addition de toutes les suites d'impulsions retardées entraîne une annulation des bandes image sauf autour des fréquences multiples de $\frac{1}{T}$, qui devient la nouvelle fréquence d'échantillonnage. C'est une application des propriétés de linéarité de la Transformée de Fourier.

Il faut maintenant faire apparaître les relations entre les transformées en Z des différentes suites d'échantillons.

La transformée en Z de la suite $s(nT)$ est définie par :

$$S(Z) = \sum_{n=-\infty}^{\infty} s(nT)\, Z^{-n} \tag{10.6}$$

Le spectre $S^M(f)$ du signal échantillonné avec la période T est obtenu en remplaçant Z par $e^{j2\pi fT}$, c'est-à-dire que :

$$S^M(f) = S(e^{j2\pi fT})$$

En décomposant la sommation dans S(Z) il vient :

$$S(Z) = \sum_{n=-\infty}^{\infty} \sum_{i=0}^{M-1} s(nMT + iT)\, Z^{-(nM+i)}$$

ou encore :

$$S(Z) = \sum_{i=0}^{M-1} Z^{-i} \sum_{n=-\infty}^{\infty} s(nMT + iT)\, Z^{-nM}$$

En posant :

$$S_i(Z^M) = \sum_{n=-\infty}^{\infty} s(nMT + iT)\, Z^{-nM}$$

il vient :

$$S(Z) = \sum_{i=0}^{M-1} Z^{-i}\, S_i(Z^M) \tag{10.7}$$

Les termes $S_i(Z^M)$ sont les transformées en Z des suites $s\left[\left(n + \frac{i}{M}\right)MT\right]$ pour $i = 0, 1, \ldots, M-1$. Le facteur Z^{-i} traduit l'entrelacement de ces suites.

Il faut maintenant exprimer $S_i(Z^M)$ en fonction de S(Z). En remplaçant dans l'équation (10.7) Z par $Z\, e^{-j2\pi m/M}$, il vient :

$$S\left(Z\, e^{-j2\pi \frac{m}{M}}\right) = \sum_{i=0}^{M-1} e^{j \frac{2\pi}{M} mi}\, Z^{-i}\, S_i(Z^M)$$

et sous forme matricielle :

$$\begin{bmatrix} S(Z) \\ S(Z\, e^{-j2\pi/M}) \\ \vdots \\ S(Z\, e^{-j2\pi(M-1)/M}) \end{bmatrix} = T_N^{-1} \begin{bmatrix} S_0(Z^M) \\ Z^{-1} S_1(Z^M) \\ \vdots \\ Z^{-(M-1)} S_{M-1}(Z^M) \end{bmatrix}$$

où T_N^{-1} est la matrice de transformation de Fourier discrète inverse du chapitre 2.

En multipliant les deux membres de cette équation par la matrice T_N, on obtient :

$$Z^{-i} S_i(Z^M) = \frac{1}{M} \sum_{m=0}^{M-1} e^{-j2\pi \frac{im}{M}}\, S\left(Ze^{-j2\pi \frac{m}{M}}\right); \quad 0 \leqslant i \leqslant (M-1) \quad (10.8)$$

ce qui correspond bien à la relation (10.5) pour les réponses en fréquence.

En posant, comme au chapitre 2 : $W = e^{-j \frac{2\pi}{M}}$, il vient :

$$S_i(Z^M) = \frac{1}{M} \sum_{m=0}^{M-1} W^{im}\, Z^i\, S(ZW^m) \qquad (10.8\ \text{bis})$$

Les expressions (10.7) et (10.8) sont des relations fondamentales en filtrage multicadence.

Les résultats obtenus, et en particulier les relations (10.3) et (10.4) sont valables pour des signaux $s(t)$ dont le spectre n'est pas limité à la fréquence $\frac{1}{2MT}$. Le repliement de bande intervient alors.

10.2 DÉCOMPOSITION D'UN FILTRE RIF PASSE-BAS

Le filtrage multicadence va d'abord être mis en évidence dans le cas des filtres RIF, où il s'introduit naturellement. En effet soit un filtre RIF passe-bas qui élimine les composantes de fréquence supérieure ou égale à la fréquence f_c dans un signal échantillonné à la fréquence f_e. Le signal filtré nécessite une fréquence d'échantillonnage égale à $2f_c$ seulement et en fait il suffit de fournir les nombres de sortie à cette cadence.

La relation qui, dans un filtre RIF d'ordre N, détermine les nombres de la suite de sortie $y(n)$, à partir de la suite des nombres d'entrée $x(n)$, s'écrit :

$$y(n) = \sum_{i=0}^{N-1} a_i x(n-i) \qquad (10.9)$$

Chaque nombre de sortie $y(n)$ est calculé à partir d'un ensemble de N nombres d'entrée, par sommation pondérée avec les coefficients a_i ($i = 0, 1, \ldots,$

N – 1). Dans ces conditions les cadences d'entrée et sortie sont indépendantes et la réduction de la cadence de sortie dans le rapport $k = \frac{f_e}{2f_c}$, supposé entier, se traduit par une réduction de la vitesse de calcul dans le même rapport.

Le même raisonnement s'applique à l'élévation de fréquence d'échantillonnage, ou interpolation. Dans ce cas la cadence en sortie est supérieure à la cadence des nombres d'entrée. Pour faire apparaître les gains en calcul il suffit simplement de considérer que les cadences sont égales en incorporant dans la suite d'entrée un nombre convenable de données nulles.

L'indépendance qui existe entre l'entrée et la sortie dans les filtres RIF peut être exploitée dans les filtres à bande passante étroite, même si les cadences d'entrée et sortie doivent être identiques, en décomposant l'opération de filtrage en deux phases [1].

– Réduction de la fréquence d'échantillonnage de la valeur f_e à une valeur intermédiaire f_0 telle que :

$$f_0 \geqslant 2f_c$$

– Élévation de la fréquence d'échantillonnage ou interpolation de f_0 à f_e.

La figure 10.3 illustre cette décomposition. Si les deux opérations sont effectuées avec un même filtre d'ordre N, le nombre de multiplications à faire par seconde M_D s'exprime par :

$$M_D = N \cdot f_0 \cdot 2 \qquad (10.10)$$

FIG. 10.3. *Filtre à Réduction-Interpolation*

Cette valeur est à comparer à la réalisation directe en un seul filtre qui conduit à la valeur M_R, telle que :

$$M_R = N \cdot f_e \qquad (10.11)$$

Par suite la décomposition est intéressante dès que le rapport $k = \frac{f_e}{2f_c}$ est supérieur à 2 c'est-à-dire dès que :

$$f_c < \frac{f_e}{4}$$

Cette approche apparaît ainsi bien adaptée aux filtres à bande passante étroite.

Il faut cependant remarquer que dans les deux cas la fonction de filtrage obtenue n'est pas rigoureusement la même et que des distorsions sont intervenues dans

la décomposition, comme le montre la figure 10.4. En effet, le sous-échantillonnage intermédiaire à la fréquence $f_0 \geqslant 2f_c$ a trois conséquences :

– Le repliement autour de la fréquence $\frac{f_0}{2}$ des composantes résiduelles de signal après filtrage aux fréquences supérieures à $\frac{f_0}{2}$. La distorsion qui en résulte est une distorsion de type harmonique; sa puissance B_R dépend de l'affaiblissement du filtre de réduction et se calcule à partir de sa fonction de transfert $H(f)$ en utilisant les résultats donnés aux chapitres précédents.

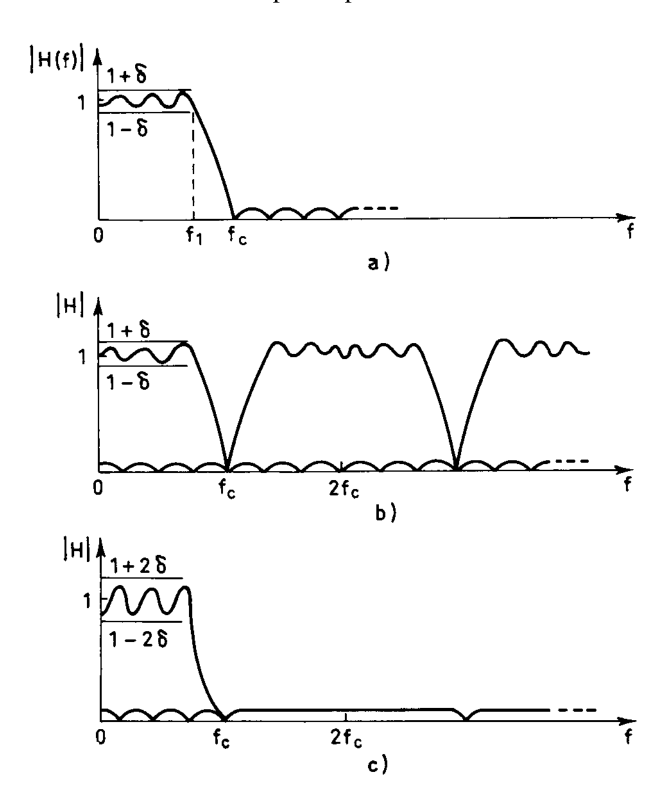


FIG. 10.4. *Réponse en fréquence du filtre multicadence*
a) Réponse du filtre direct
b) Réponse en sortie du filtre de réduction
c) Réponse du filtre multicadence

Par exemple, si le signal d'entrée a une distribution spectrale uniforme et une puissance unitaire, la puissance totale B_T du signal replié s'écrit :

$$B_T = \frac{1}{f_e} \int_{\frac{f_0}{2}}^{f_e - \frac{f_0}{2}} |H(f)|^2 \, df \qquad (10.12)$$

En désignant par f_1 la limite de la bande passante du filtre, un majorant de B_T est fourni par la relation :

$$B_T < \sum_{i=0}^{N-1} |a_i|^2 - \frac{2f_1}{f_e}$$

La distorsion peut être supposée à distribution spectrale uniforme et seule la puissance en bande passante est à considérer, il vient dans ce cas :

$$B_R < \frac{2f_1}{f_0}\left[\sum_{i=0}^{N-1} |a_i|^2 - \frac{2f_1}{f_2}\right] \qquad (10.13)$$

Il faut tenir compte de cette dégradation du signal dans le calcul du filtre de réduction [2].

– La périodicité en fréquence de la réponse du filtre de réduction avec la période f_0 introduit une distorsion dont la puissance B_i est fonction de l'affaiblissement du filtre interpolateur.

Si ce filtre est le même que le filtre de réduction, avec les mêmes hypothèses, on obtient comme précédemment :

$$B_i = \frac{1}{f_e}\int_{\frac{f_0}{2}}^{f_e-\frac{f_0}{2}} |H(f)|^2\, df$$

Cette distorsion, extérieure à la bande passante, peut être gênante pour l'addition d'autres signaux au signal filtré.

– La mise en cascade de deux filtres augmente les ondulations en bande passante. Par exemple ces ondulations sont doublées si le même filtre est utilisé dans les deux opérations.

Finalement les sous-ensembles du filtre multicadence doivent être conçus pour que l'ensemble satisfasse aux caractéristiques globales imposées au filtre.

Le schéma de la figure 10.3 se simplifie si les fréquences d'échantillonnage du signal avant et après filtrage peuvent être différentes. Le principe exposé peut aussi s'appliquer aux filtres passe-haut et passe-bande, moyennant l'introduction d'étages de modulation et démodulation par exemple.

Le principe de décomposition peut être étendu au sous-ensemble réducteur de fréquence d'échantillonnage et au sous-ensemble interpolateur, ce qui apporte un gain supplémentaire. Un filtre élémentaire particulièrement efficace pour réaliser ces sous-ensembles est le filtre RIF demi-bande.

10.3 LE FILTRE RIF DEMI-BANDE

Le filtre RIF demi-bande a été présenté au paragraphe 5.8 ; c'est un filtre à phase linéaire dont la réponse en fréquence $H(f)$ prend la valeur $\frac{1}{2}$ à la fréquence $\frac{f_e}{4}$

et est antisymétrique par rapport à ce point, c'est-à-dire que la fonction $H(f)$ vérifie les relations :

$$H\left(\frac{f_e}{4}\right) = 0{,}5\,; \quad H\left(\frac{f_e}{4}+f\right) = 1 - H\left(\frac{f_e}{4}-f\right) \tag{10.14}$$

Pour un nombre de coefficients $N = 4M + 1$, on a :

$$H(f) = e^{-j2\pi 2Mf}\cdot\frac{1}{2}\cdot\left[1 + 2\sum_{i=1}^{M} h_{2i-1}\cos\left[2\pi(2i-1)f\right]\right] \tag{10.15}$$

Les coefficients h_i sont nuls pour i pair sauf h_0. La figure 10.5 illustre les caractéristiques de ce filtre. Le gabarit est défini par l'ondulation en bandes passante et affaiblie δ et par la largeur Δf de la bande de transition ; ces paramètres étant donnés, le nombre de coefficients N peut être estimé par les formules données au paragraphe 5.6. Dans ce cas particulier, ces formules peuvent être simplifiées et d'après (5.32) il vient :

$$N \simeq \frac{2}{3}\log\left(\frac{1}{10\,\delta^2}\right)\frac{f_e}{\Delta f}$$

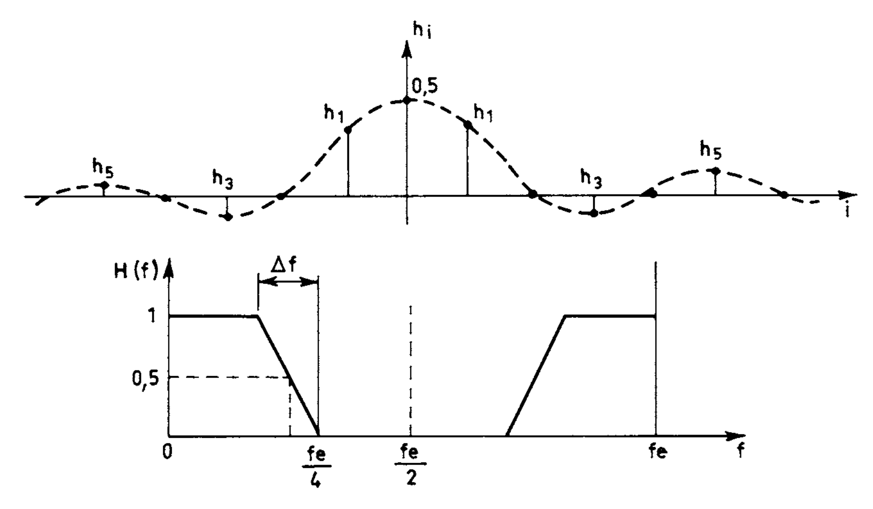


FIG. 10.5. *Filtre RIF demi-bande*

En considérant l'affaiblissement A_f tel que :

$$A_f = 10\log\left(\frac{1}{\delta^2}\right)$$

et compte tenu du rôle particulier que joue la fréquence $\frac{f_e}{4}$ dans ce type de filtre, on peut écrire :

$$M \simeq \frac{2}{3}\left[\frac{A_f}{10} - 1\right]\cdot\frac{f_e}{4\Delta f}$$

D'où la relation très simple suivante entre l'affaiblissement exprimé en décibels et la bande de transition, pour un nombre donné de coefficients :

$$A_f = 10 + 15 \,.\, M \,.\left(\frac{4\Delta f}{fe} \right) \qquad (10.16)$$

En fait c'est une relation approchée valable dès que M dépasse quelques unités.

Les coefficients se calculent par le programme général de calcul des filtres RIF, il suffit d'introduire les données correspondantes.

La relation entre les suites d'entrée et sortie s'écrit :

$$y(n) = \frac{1}{2}\left[x(n-2M) + \sum_{i=1}^{M} h_{2i-1}\left[x(n-2M+2i-1) + x(n-2M-2i+1) \right] \right] \qquad (10.17)$$

et le nombre de multiplications à faire pour chaque élément de la suite de sortie $y(n)$ est égal à M. On peut remarquer que ces opérations ne portent que sur les éléments de la suite d'entrée d'indice impair. Il s'ensuit que si un tel filtre est utilisé pour réduire la fréquence d'échantillonnage de la valeur f_e à la valeur $f_0 = \frac{f_e}{2}$, le nombre de multiplications à faire par seconde est égal à $M \,.\, f_0$. Il en est de même pour élever la fréquence d'échantillonnage de f_0 à $f_e = 2f_0$, l'opération consistant alors simplement à calculer un échantillon entre deux échantillons de la suite d'entrée.

Finalement le nombre de multiplications à faire par seconde dans un filtre demi-bande avec changement de fréquence d'échantillonnage s'écrit :

$$M_R = \left[\frac{2}{3} \log\left(\frac{1}{10 \,.\, \delta^2} \right) \cdot \frac{fe}{\Delta f} \cdot \frac{1}{4} \right] \frac{fe}{2} \qquad (10.18)$$

Le nombre de mémoires de données dans une réduction de fréquence d'échantillonnage est le nombre nécessaire pour stocker les données sur lesquelles porte la sommation pondérée, c'est-à-dire : $MM_D = 2M$. Dans une interpolation il faut non seulement calculer la somme pondérée, mais entrelacer les résultats avec la suite d'entrée retardée de M périodes ; alors on a :

$$MM_D = 3M.$$

Dans les deux cas le nombre de mémoires de coefficients est égal à

$$MM_C = M.$$

Exemple :

Un ensemble de filtres demi-bande ayant des caractéristiques utiles dans les applications est fourni par le tableau 10.1 qui donne les coefficients quantifiés, l'échelon de quantification étant pris comme unité [3].

La réponse en fréquence se calcule simplement par application de la relation (10.15). Les filtres F4, F6, F8 et F9 correspondent à une valeur du paramètre $4\,\frac{\Delta f}{f_e}$

égale à l'unité, avec des ondulations de 37, 50, 67 et 79 dB respectivement. Les filtres F2, F3 et F5 ont des réponses monotones.

Les avantages de la structure de filtre particulière décrite dans ce paragraphe peuvent être utilisés pour le filtrage multicadence général.

Tableau 10.1. – Filtres RIF demi-bande

Filtre	h_0	h_1	h_3	h_5	h_7	h_9
F1	1	1				
F2	2	1				
F3	16	9	– 1			
F4	32	19	– 3			
F5	256	180	– 25	3		
F6	346	208	– 44	9		
F7	512	302	– 53	7		
F8	802	490	– 116	33	– 6	
F9	8192	5042	– 1277	429	– 116	18

10.4 DÉCOMPOSITION AVEC FILTRES DEMI-BANDE

L'exploitation des particularités du filtre élémentaire demi-bande conduit au schéma de filtre multicadence donné à la figure 10.6. La fréquence intermédiaire f_0 est avec la fréquence d'échantillonnage f_e dans un rapport qui est une puissance de deux :

$$f_e = 2^P \cdot f_0 \tag{10.19}$$

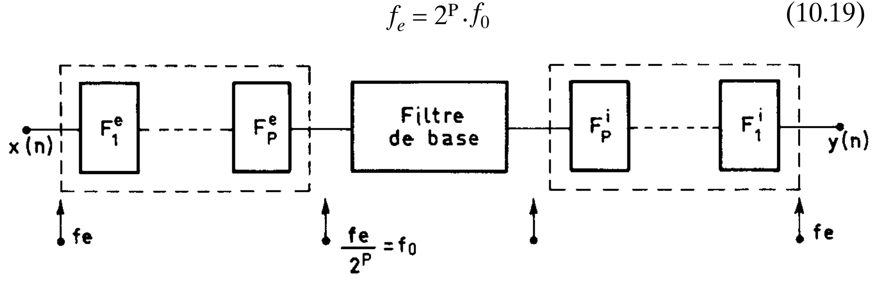


Fig. 10.6. *Décomposition avec filtres demi-bande*

La réduction et l'élévation de fréquence d'échantillonnage sont faites par une mise en cascade de P filtres demi-bande. L'ensemble comporte un filtre de base fonctionnant à la fréquence f_0 et encadré de deux cascades de filtres demi-bande.

Le filtre passe-bas global est spécifié par la donnée des paramètres suivants :

- Ondulation en bande passante : δ_1
- Ondulation en bande affaiblie : δ_2
- Largeur de bande de transition : Δf
- Fin de bande passante : f_1
- Début de bande affaiblie : f_2

Pour calculer les filtres demi-bande il faut définir leurs spécifications. L'ondulation en bande passante est supposée partagée entre les filtres demi-bande et le filtre de base. D'autre part, chaque filtre doit avoir une ondulation en bande affaiblie meilleure que δ_2. Il en résulte que pour chaque filtre demi-bande, l'ondulation δ_0 est donnée par :

$$\delta_0 = \min\left\{\frac{\delta_1}{4P}, \delta_2\right\} \tag{10.20}$$

Le premier filtre demi-bande F_1^e de la cascade peut être déterminé si sa bande de transition Δf_1 est fixée (fig. 10.7).

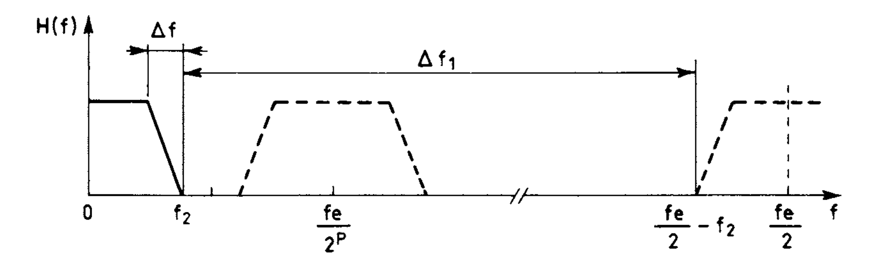


Fig. 10.7. *Bande de transition du premier filtre demi-bande*

Pour fixer Δf_1 il faut considérer que le rôle du premier filtre est d'éliminer les composantes du signal qui pourraient se replier dans la bande utile après division par deux de la fréquence d'échantillonnage. Il vient :

$$\Delta f_1 = \frac{f_e}{2} - 2f_2$$

D'après la relation (10.18) le nombre de multiplications M_{C_1} à faire dans le premier filtre s'écrit :

$$M_{C_1} = \frac{2}{3}\log\left(\frac{1}{10\,\delta_0^2}\right)\cdot\frac{f_e}{\Delta f_1}\cdot\frac{1}{4}$$

En introduisant le paramètre k, qui traduit l'écart entre f_2 et $\frac{1}{2}\,\frac{f_e}{2^P}$, et tel que :

$$k = \frac{f_2}{f_e}\,.2^{P+1}$$

et en posant :

$$D(\delta_0) = \frac{2}{3}\log\left(\frac{1}{10\,\delta_0^2}\right)$$

il vient :

$$M_{C_1} = D(\delta_0)\,.\left[\frac{1}{2} - \frac{k}{2^P}\right]^{-1}\cdot\frac{1}{4}$$

La bande de transition du i^e filtre de la cascade s'écrit :

$$\Delta f_i = \frac{f_e}{2^i} - 2f_2 = f_e\left[\frac{1}{2^i} - \frac{k}{2^P}\right]$$

Comme la fréquence d'échantillonnage d'entrée de ce filtre a la valeur $\frac{f_e}{2^{i-1}}$, le nombre de multiplications M_{C_i} à faire s'exprime par :

$$M_{C_i} = D(\delta_0) \cdot \frac{1}{2^{i-1}} \cdot \left(\frac{1}{2^i} - \frac{k}{2^P}\right)^{-1} \cdot \frac{1}{4}$$

La fréquence d'échantillonnage en sortie du i^e filtre étant $\frac{f_e}{2^i}$, le nombre total M_C de multiplications à faire dans la cascade de P filtres résulte de la sommation :

$$M_C = \sum_{i=1}^{P} D(\delta_0) \cdot \frac{1}{2^{i-1}} \cdot \left(\frac{1}{2^i} - \frac{k}{2^P}\right)^{-1} \cdot \frac{1}{4} \cdot \frac{f_e}{2^i}$$

Soit :

$$M_C = \frac{1}{2} \cdot D(\delta_0) \cdot f_e \cdot \sum_{i=1}^{P} \frac{1}{2^i}\left[1 - \frac{k}{2^{P-i}}\right]^{-1}$$

La fonction $S(k)$ telle que :

$$S(k) = \sum_{i=1}^{P} \frac{1}{2^i}\left[1 - \frac{k}{2^{P-i}}\right]^{-1} \quad (10.21)$$

prend des valeurs voisines de l'unité sauf pour les faibles valeurs de P et quand k s'approche de l'unité, ce qu'il convient de chercher à éviter en réduisant par exemple d'une unité le nombre de filtres de la cascade. Par suite on peut donner l'expression approchée simple :

$$M_C \simeq \frac{1}{3} \cdot \log\left(\frac{1}{10\,\delta_0^2}\right) \cdot f_e \quad (10.22)$$

Le nombre de mémoires pour les coefficients est égal dans chaque filtre au nombre de multiplications M_{C_i}. Pour l'exprimer il suffit d'une sommation. Il vient :

$$MM_{CC} = \sum_{i=1}^{P} M_{C_i} = \frac{1}{2} \cdot D(\delta_0) \cdot \sum_{i=1}^{P}\left[1 - \frac{k}{2^{P-i}}\right]^{-1} \quad (10.23)$$

Cette valeur peut être approchée par :

$$MM_{CC} \simeq \frac{1}{2} \cdot D(\delta_0) \cdot \left[P + k\left[\frac{1}{1-k} + \frac{1}{1-2k}\right]\right] \quad (10.24)$$

Le nombre de mémoires de données est égal au double de la valeur obtenue pour les coefficients, avec des choix de structures convenables, comme on le verra ultérieurement.

Pour estimer le volume de calculs à faire par seconde dans le filtre complet il faut déterminer l'ordre N du filtre de base :

$$N \simeq D\left(\frac{\delta_1}{2}, \delta_2\right) \cdot \frac{1}{\Delta f} \cdot \frac{f_e}{2^P}$$

avec :

$$D\left(\frac{\delta_1}{2}, \delta_2\right) = \frac{2}{3} . \log\left(\frac{1}{5\,\delta_1\,\delta_2}\right)$$

Les valeurs des paramètres de complexité pour le filtre complet représenté à la figure 10.6 sont finalement les suivantes :

– Nombre de multiplications par seconde :

$$M_R = f_e\left[D(\delta_0) + \frac{1}{2^{2P+1}} \cdot \frac{f_e}{\Delta f} . D\left(\frac{\delta_1}{2}, \delta_2\right)\right] \tag{10.25}$$

– Nombre de mémoires de données :

$$MM_D = D\left(\frac{\delta_1}{2}, \delta_2\right) \cdot \frac{f_e}{\Delta f} \cdot \frac{1}{2^P} + 2D(\delta_0)\left[P + k\left(\frac{1}{1-k} + \frac{1}{2-k}\right)\right] \tag{10.26}$$

– Nombre de mémoires de coefficients :

$$MM_C = D\left(\frac{\delta_1}{2}, \delta_2\right) \cdot \frac{f_e}{\Delta f} \cdot \frac{1}{2^{P+1}} + \frac{1}{2} D(\delta_0)\left[P + k\left(\frac{1}{1-k} + \frac{1}{2-k}\right)\right] \tag{10.27}$$

Cette estimation est basée sur l'hypothèse que les deux cascades de filtres demi-bande sont identiques. Les relations (10.25-26-27) sont utiles dans les projets pour évaluer l'intérêt dans chaque cas du filtrage multicadence. Les deux exemples ci-dessous illustrent leur application.

Exemple 1 :

Soit un filtre passe-bas étroit défini par les valeurs de paramètres suivantes :

$$f_e = 1; \quad f_2 = 0{,}05; \quad \Delta f = 0{,}025; \quad \delta_1 = 0{,}01; \quad \delta_2 = 0{,}001.$$

Les paramètres ont les valeurs :

$$P = 3; \quad \delta_0 = \min\left\{\frac{0{,}01}{12}\,; 0{,}001\right\} = 0{,}00083; \quad D(\delta_0) = 3{,}3$$

$$D\left(\frac{\delta_1}{2}, \delta_2\right) = 2{,}76; \quad k = 0{,}8; \quad S(k) = 1{,}6$$

Il vient :

$$M_R = 6{,}2; \quad MM_D = 65; \quad MM_c = 20$$

Une réalisation directe conduit à un filtre d'ordre N = 110, ce qui correspond aux valeurs :

$$M_R = 55; \quad MM_D = 110; \quad MM_c = 55$$

Exemple 2 :

Soit un filtre très étroit tel que :

$$f_e = 1 ; \quad f_2 = 0{,}005 ; \quad \Delta f = 0{,}00025 ; \quad \delta_1 = 0{,}001 ; \quad \delta_2 = 0{,}0001$$

Les paramètres ont les valeurs :

$$P = 6 ; \quad \delta_0 = \min\left\{\frac{0{,}001}{24} ; 0{,}0001\right\} = 0{,}0000416$$

$$D(\delta_0) = 5{,}38 ; \quad D\left(\frac{\delta_1}{2}, \delta_2\right) = 4{,}72 ; \quad k = 0{,}64 ; \quad S(k) = 1{,}07.$$

Il vient :

$$M_R = 7{,}76 ; \quad MM_D = 350 ; \quad MM_C = 152$$

Une réalisation directe conduirait à un filtre dont l'ordre N serait de plusieurs milliers et qui ne peut être envisagé pratiquement.

Ces exemples font bien apparaître l'avantage des techniques de filtrage multicadence. Il est intéressant également de faire une comparaison avec les filtres RII. Cette comparaison va être menée sur un filtre satisfaisant au gabarit suivant :

$$\delta_1 = \delta_2 = 10^{-2}$$

$$\Delta f = \frac{f_1 + f_2}{2} . 10^{-1}$$

et pour lequel on fait varier la position de la bande de transition, c'est-à-dire le rapport :

$$\frac{2 f_e}{f_1 + f_2}$$

Le filtre RII est supposé être de type elliptique et réalisé en cellules du second ordre demandant 4 multiplications chacune. Supposant $f_e = 1$ les nombres de multiplications à faire par seconde et de mémoires de données sont reportés au tableau 10.2.

Tableau 10.2. – Comparaison des complexités des filtres RIF, Multicadence et RII

$\frac{2f_e}{f_1+f_2}$	Multiplications			Mémoires		
	RIF	Multicadence	RII	RIF	Multicadence	RII
2	45	23	15	90	90	7
3	65	19	15	130	80	7
5	110	12	15	220	95	7
10	220	8	15	440	105	7

Les résultats montrent que le filtrage multicadence est nettement plus avantageux que le filtre RIF aussi bien en mémoires de données qu'en multiplications. Par contre, le filtre RII, étant à phase minimale, est celui qui nécessite le minimum de mémoires.

On peut observer que si le filtre de base est un filtre d'ordre élevé et si la linéarité en phase n'est pas indispensable, il est alors avantageux de remplacer le filtre de base, qui fonctionne à la cadence f_0 en entrée et sortie, par un filtre RII. Une réduction substantielle du volume de calcul peut ainsi être obtenue.

Le filtrage multicadence utilisant des cascades de filtres demi-bande s'avère ainsi être avantageux pour la quantité de calculs et de mémoires, mais la réalisation d'un filtre par une série de sous-ensembles fonctionnant à des cadences différentes est susceptible de compliquer l'enchaînement des calculs et de réagir sur la complexité de l'unité de commande des systèmes ou sur la taille des mémoires de programme des calculateurs de traitement.

D'autres structures que les filtres demi-bandes peuvent être envisagées en filtrage multicadence. D'abord on peut chercher à réduire le nombre d'étages, en utilisant des variations de fréquences d'échantillonnage par des facteurs supérieurs à deux. Les techniques de choix des facteurs les plus intéressants sont données dans la référence [4], le volume de calcul nécessaire est plus élevé qu'avec les filtres demi-bande mais l'enchaînement est plus simple, ce qui peut être appréciable si un calculateur universel est envisagé.

Le filtrage multicadence, tel qu'il a été présenté dans les paragraphes précédents, repose sur une décomposition des filtres avec introduction de sous-ensembles de type non récursif. Dans un but de généralisation, la question se pose de savoir s'il est possible d'obtenir des gains en calcul avec des sous-ensembles récursifs. En fait, il est important, afin de pouvoir optimiser le traitement multicadence dans des cas plus généraux que la réalisation d'un seul filtre, de mettre en évidence la fonction de déphaseur.

10.5 FILTRAGE PAR RÉSEAU POLYPHASÉ

La mise en évidence des relations de phase entre différents échantillonnages d'un même signal a été faite au paragraphe 10.1. Les résultats vont être utilisés pour analyser sous cet aspect le filtrage multicadence [5].

Soit une réduction de la fréquence d'échantillonnage f_e par un facteur N ; la suite d'entrée $x(n)$ a pour transformée en Z la fonction X(Z) ; la transformée de Fourier de cette suite est obtenue en remplaçant Z par $e^{j2\pi \frac{f}{f_e}}$, soit $X(e^{j2\pi \frac{f}{f_e}})$. La suite de sortie $y(\mathrm{N}n)$ échantillonnée à la fréquence $\frac{f_e}{\mathrm{N}}$ a pour transformée en Z une fonction de la variable Z^N : $Y(Z^N)$. Par suite si des circuits déphaseurs interviennent dans cette opération leur fonction de transfert est aussi fonction de la variable Z^N et peut être calculée à partir de la fonction de filtrage global.

La structure de déphaseur va être mise en évidence d'abord dans l'élément particulièrement simple que constitue le filtre RIF demi-bande, qui est défini par la relation (10.17).

Cette relation peut être réécrite comme suit :

$$y(n) = \frac{1}{2}\left[x(n-2M) + \sum_{i=1}^{2M} a_i x(n-2i+1)\right] \qquad (10.28)$$

avec :

$$a_i = h_{(2M-2i+1)} = a_{2M-i+1} \quad \text{pour} \quad 1 \leqslant i \leqslant M$$

La fonction de transfert en Z correspondante a pour expression :

$$H(Z) = \frac{1}{2}\left[Z^{-2M} + Z^{-1}.\sum_{i=0}^{2M-1} a_{i+1} Z^{-2i}\right] \qquad (10.29)$$

ou encore :

$$H(Z) = \frac{1}{2}\left[H_0(Z^2) + Z^{-1} H_1(Z^2)\right] \qquad (10.30)$$

La réponse en fréquence correspondante s'écrit :

$$H(f) = \frac{1}{2}\left[e^{-j2\pi\frac{f}{f_e}2M} + e^{-j2\pi\frac{f}{f_e}}.H_1(f)\right] \qquad (10.31)$$

La fonction $H_0(f)$ est une caractéristique de déphaseur pur linéaire, puisqu'elle correspond à un retard.

En raison de la symétrie des coefficients, la fonction $H_1(f)$ est également à phase linéaire. Cette partie du filtre fonctionnant à la cadence $\frac{f_e}{2}$, $H_1(f)$ présente la périodicité $\frac{f_e}{2}$. Comme le nombre de coefficients est pair, d'après les résultats du paragraphe V.2 on peut écrire :

$$H_1(f) = e^{-j2\pi f\left(M-\frac{1}{2}\right)\frac{2}{f_e}}.R(f)$$

ou encore, en explicitant la phase.

$$H_1(f) = e^{-j2\pi\left(\frac{f}{f_e}\right)2M}.e^{-j\varphi(f)}.R(f) \qquad (10.32)$$

La fonction $\varphi(f)$ est linéaire et présente la périodicité $\frac{f_e}{2}$. Par suite, elle s'exprime par :

$$\varphi(f) = \pi\left(\left[\frac{2f}{f_e} + \frac{1}{2}\right] - \frac{2f}{f_e}\right) \qquad (10.33)$$

où $[x]$ représente le plus grand entier contenu dans x.

Pour l'amplitude, on peut reprendre la relation (10.8), avec la symétrie de $H(f)$, ce qui conduit à :

$$e^{-j2\pi \frac{f}{f_e}} H_1(f) = H(f) - H\left(\frac{f_2}{2} - f\right)$$

En introduisant la propriété d'antisymétrie par rapport à la fréquence $\frac{f_e}{4}$ démontrée au paragraphe (10.4), il vient :

$$|H_1(f)| = 2|H(f)| - 1; \quad 0 \leqslant f \leqslant \frac{f_e}{4}$$

Les ondulations sont doubles de celles de $H(f)$ et la fonction s'annule pour $f = \frac{f_e}{4}$, fréquence qui correspond au changement de phase de π.

La figure 10.8 représente les fonctions $|H_1(f)|$ et $\varphi(f)$ qui caractérisent le filtre.

La phase $\Phi(f)$ telle que :

$$\Phi(f) = \varphi(f) + 2\pi \frac{f}{f_e} \tag{10.34}$$

est constante et prend les valeurs 0 ou π.

Finalement le circuit dont la réponse en fréquence est la fonction $e^{-j2\pi \frac{f}{f_e}}$. $H_1(f)$ approche un déphaseur pur dans les bandes utiles, c'est-à-dire en bandes passante et affaiblie ; le nombre de ses coefficients et sa complexité dépendent du degré d'approximation, c'est-à-dire de la bande de transition Δf et de l'ondulation en amplitude δ. Ces résultats sont à rapprocher de ceux du paragraphe 9.3.

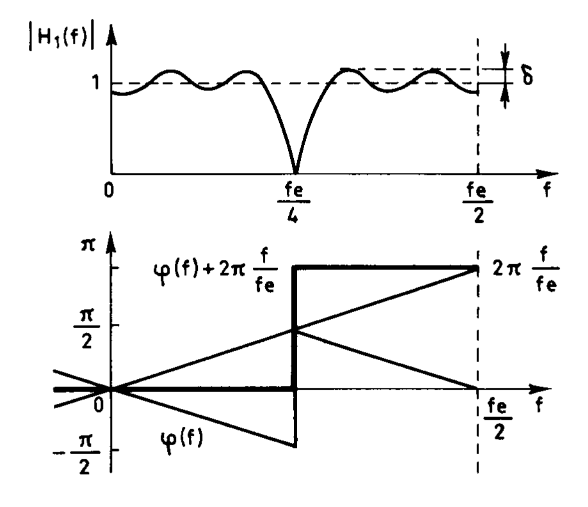


FIG. 10.8. *Fonction $H_1(f)$ du filtre demi-bande*

Le filtre demi-bande se présente comme un réseau à 2 branches, selon la figure 10.9. La réponse globale correspond à la somme des réponses des deux branches comme le montre la figure 10.11.

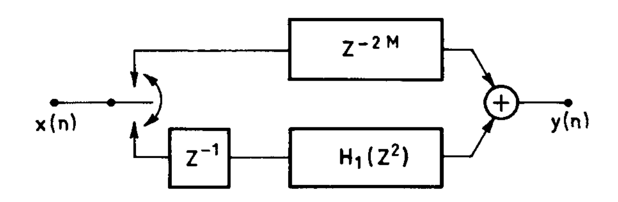


FIG. 10.9. *Filtre demi-bande avec déphaseurs*

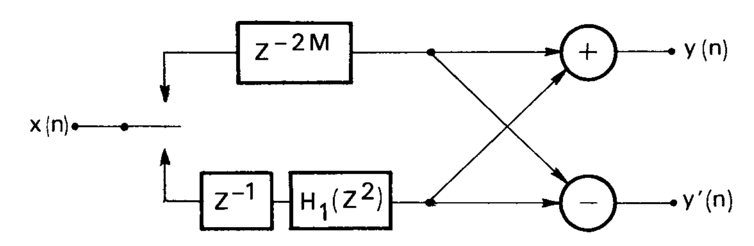


FIG. 10.10. *Banc de deux filtres*

Si l'on change le signe de la fonction $H_1(f)$, alors, comme le montre la figure 10.11, un filtre passe-haut est obtenu. Ainsi un ensemble de deux filtres est obtenu avec les mêmes calculs. Soient $B_0(Z)$ et $B_1(Z)$ les fonctions de transfert de ces filtres ; le système correspondant, représenté sur la figure 10.10, est caractérisé par l'équation matricielle suivante :

$$\begin{bmatrix} B_0(Z) \\ B_1(Z) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} Z^{-2M} \\ Z^{-1}H_1(Z^2) \end{bmatrix} \qquad (10.35)$$

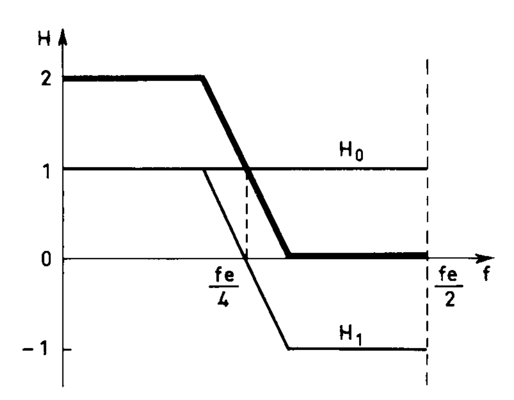


FIG. 10.11. *Réponse en amplitude du filtre demi-bande*

On reconnaît la matrice de la Transformée de Fourier d'ordre 2. La généralisation de ce résultat à un banc de N filtres fait intervenir une Transformée de Fourier d'ordre N.

On peut aussi remarquer que, en changeant dans $H_1(f)$ le signe d'un coefficient sur deux, la réponse en fréquence se trouve décalée de $\frac{f_e}{4}$, et on obtient le filtre de quadrature du paragraphe 9.3.

Les résultats obtenus pour le filtre demi-bande se généralisent facilement à un filtre RIF utilisé pour réduire ou élever la fréquence d'échantillonnage par le facteur N. Soit $H(Z)$ la fonction de transfert en Z d'un tel filtre. En supposant qu'il possède KN coefficients, on peut écrire :

$$H(Z) = \sum_{i=1}^{KN} a_i Z^{-i} = \sum_{n=0}^{N-1} Z^{-n} H_n(Z^n) \tag{10.36}$$

avec :

$$H_n(Z^N) = a_{kN+n}(Z^{-N})^k$$

Ce filtre peut être réalisé par un réseau à N branches, suivant la figure 10.12, appelé réseau polyphasé, car chaque branche a une réponse en fréquence qui approche celle d'un déphaseur pur. Les déphasages sont constants par plage de fréquence et multiples entiers de $\frac{2\pi}{N}$. Quand il y a changement de fréquence d'échantillonnage dans un rapport N, les circuits des différentes branches du réseau opèrent à la fréquence $\frac{f_e}{N}$.

FIG. 10.12. *Filtrage par réseau polyphasé*

Les filtres à Réponse Impulsionnelle Infinie ayant des sélectivités plus grandes que les filtres à Réponse Impulsionnelle Finie, il est intéressant d'étudier les filtres multicadence à éléments récursifs.

10.6 FILTRAGE MULTICADENCE À ÉLÉMENTS RII

La technique de base pour le calcul d'un filtre multicadence à éléments RII consiste à faire le même type de décomposition de la fonction de transfert du filtre RII global H(Z) que celui que fournit la relation (10.36). La fonction H(Z) est supposée être une fraction rationnelle où le dénominateur et le numérateur sont de même degré.

Une telle décomposition est obtenue en faisant apparaître les pôles de H(Z) :

$$H(Z) = a_0 \frac{\prod_{k=1}^{K} (Z - Z_k)}{\prod_{k=1}^{K} (Z - P_k)} \qquad (10.37)$$

En utilisant l'identité :

$$Z^N - P_k^N = (Z - P_k)\,(Z^{N-1} + Z^{N-2} P_k + \ldots + P_k^{N-1}) \qquad (10.38)$$

On obtient :

$$H(Z) = a_0 \frac{\prod_{k=1}^{K} (Z - Z_k)\,(Z^{N-1} + P_k Z^{N-2} + \ldots + P_k^{N-1})}{\prod_{k=1}^{K} (Z^N - P_k^N)}$$

qui s'écrit sous une autre forme :

$$H(Z) = \frac{\sum_{i=0}^{KN} a_i Z^{-i}}{1 + \prod_{k=1}^{K} b_k Z^{-Nk}}$$

Il vient alors :

$$H_n(Z^N) = \frac{\sum_{k=0}^{K} a_{kN+n} Z^{-Nk}}{1 + \sum_{k=1}^{K} b_k Z^{-Nk}} \qquad (10.39)$$

ou encore :

$$H_n(Z^N) = \frac{N_n(Z^N)}{D(Z^N)} \qquad (10.40)$$

Toutes les branches du réseau polyphasé sont déterminées. Elles ont toutes la même partie récursive et se distinguent par la partie non récursive comme le montre la relation (10.40). Dans le principe, la différence par rapport au paragraphe précédent est que les déphaseurs RII obtenus, pris individuellement, ne sont pas à phase linéaire.

La structure pour réaliser le filtre multicadence sous cette forme est donnée à la figure 10.13 dans le cas de l'augmentation de fréquence d'échantillonnage.

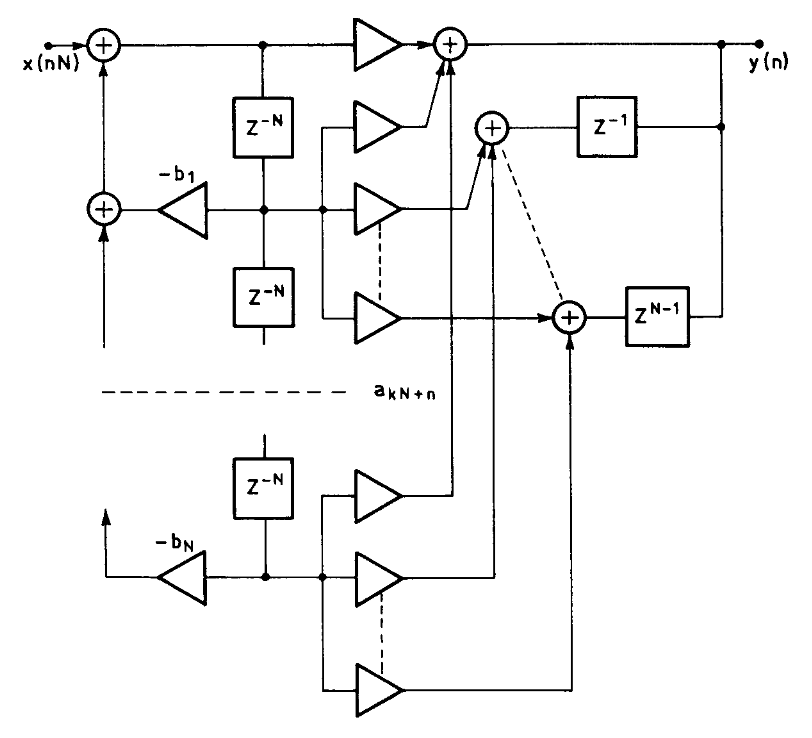


FIG. 10.13. *Augmentation de fréquence d'échantillonnage par une structure récursive*

Tous les calculs se font à la cadence $\frac{1}{NT}$, mais on peut observer qu'il y a à faire un nombre de multiplications égal à KN + K + 1. Une réalisation directe du filtre RII demande dans les mêmes conditions (2K + 1).N multiplications. Par suite la décomposition apporte un gain qui est seulement de l'ordre du facteur 2. Le véritable intérêt de cette décomposition est pour les bancs de filtres.

Dans la procédure ci-dessus, il faut toutefois remarquer que la partie récursive du circuit de la figure 10.13 correspond à des pôles élevés à la puissance N, P_k^N. Cette transformation est représentée sur la figure 10.14 pour K = 8.

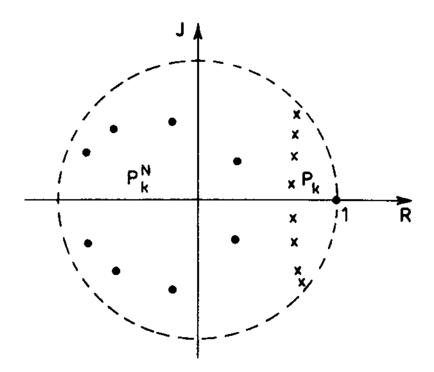


FIG. 10.14. *Pôles du réseau polyphasé*

Elle montre que les pôles se dispersent à l'intérieur du cercle unité et s'éloignent de ce cercle [5]. En effet si l'on a :

$$|P_k| = 1 - \varepsilon \quad \text{avec } \varepsilon \text{ très petit et positif}$$

il vient après élévation à la puissance N :

$$|P_k^N| \simeq 1 - N\varepsilon$$

C'est un élément très favorable pour l'incidence sur la fonction de transfert de la représentation des coefficients du dénominateur, comme le montre le paragraphe 7.7 et en particulier la relation (7.66). En fait, la structure est réalisable sous forme directe alors que le filtre initial ne l'est pas.

Les incidences de la limitation du nombre de bits des coefficients s'analysent comme au chapitre 7. Il en est de même de l'évaluation de la puissance du bruit de calcul.

10.7 BANC DE FILTRES PAR RÉSEAU POLYPHASÉ ET TFD

Un calculateur de Transformée de Fourier Discrète constitue un banc de filtres (paragraphe 2.4) qui sont bien adaptés au filtrage multicadence.

Cependant il faut remarquer que les filtres ainsi réalisés présentent des recouvrements importants. Pour améliorer la discrimination entre les composantes du signal, on effectue une pondération des nombres avant application de la TFD. Les coefficients de pondération sont les échantillons de fonctions dites fenêtres d'analyse spectrale. La réalisation de bancs de filtres par réseau polyphasé et TFD constitue en fait la généralisation des fenêtres d'analyse spectrale [6, 7].

Soit à réaliser un banc de N filtres qui couvrent la bande $[0, f_e]$ et sont obtenus par translation en fréquence d'un filtre de base ou filtre prototype, de la valeur $m.\dfrac{f_e}{N}$ avec $1 \leqslant m \leqslant N-1$.

Si H(Z) est la fonction de transfert en Z du filtre, une translation en fréquence de $m.\dfrac{f_e}{N}$ se traduit par un changement de variable de Z en $Z.e^{j2\pi\frac{m}{N}}$; c'est-à-dire que le filtre d'indice m a pour fonction de transfert $B_m(Z)$ telle que :

$$B_m(Z) = H(Z.e^{j2\pi\frac{m}{N}})$$

En appliquant la décomposition de H(Z) introduite dans les paragraphes précédents, il vient :

$$B_m(Z) = \sum_{n=0}^{N-1} Z^{-n}.e^{-j2\pi\frac{mn}{N}}.H_n(Z^N)$$

En tenant compte du fait que les fonctions $H_n(Z^N)$ sont les mêmes pour tous les filtres, une factorisation peut intervenir, conduisant à l'équation matricielle suivante :

$$\begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_{N-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & W & W^2 & \dots & W^{N-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & W^{N-1} & W^{2(N-1)} & \dots & W^{(N-1)^2} \end{bmatrix} \begin{bmatrix} H_0(Z^N) \\ Z^{-1}H_1(Z^N) \\ \vdots \\ Z^{-(N-1)}H_N(Z^N) \end{bmatrix} \tag{10.41}$$

où $W = e^{-j\frac{2\pi}{N}}$.

La matrice carrée est la matrice de la TFD. Le banc de filtres est réalisé par mise en cascade du réseau polyphasé de la figure 10.12 et d'un calculateur de TFD.

Le fonctionnement de ce dispositif est illustré par la figure 10.15, qui montre les déphasages qui interviennent aux différents points du système pour aboutir à ne conserver que le signal compris dans la bande $\left[\frac{1}{2NT}, \frac{3}{2NT}\right]$, dans le cas où $N = 4$.

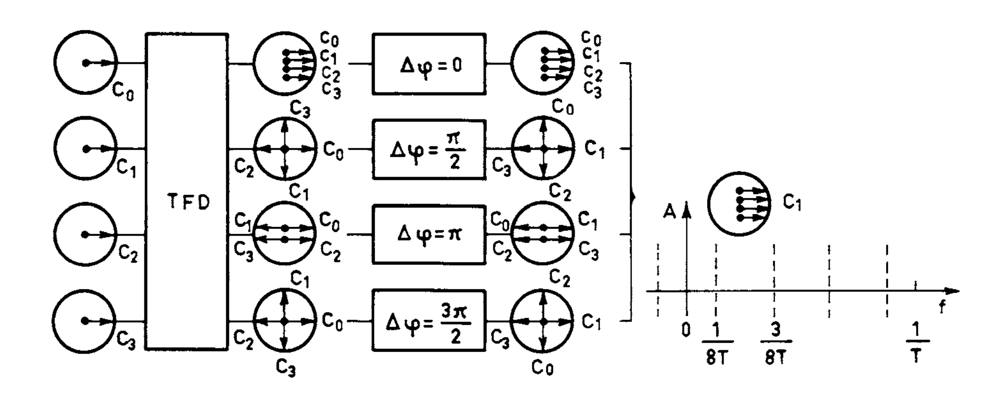


Fig. 10.15. *Déphasages dans un banc de 4 filtres*

Le réseau polyphasé a pour effet de corriger, dans la partie utile de la bande élémentaire $\left[\frac{1}{2NT}, \frac{3}{2NT}\right]$, l'effet de l'entrelacement des nombres en sortie du calculateur de TFD, ce qui permet d'éviter le recouvrement entre les filtres et conduit à la fonction de filtrage de la figure 10.16. Cette fonction ne dépend que du filtre de base H(Z) qui peut être soit de type RIF soit de type RII, et qui peut être spécifié pour que les filtres du banc n'aient aucun recouvrement, ou au contraire aient par exemple un point d'intersection à 6 dB, ou 3 dB comme au chapitre suivant.

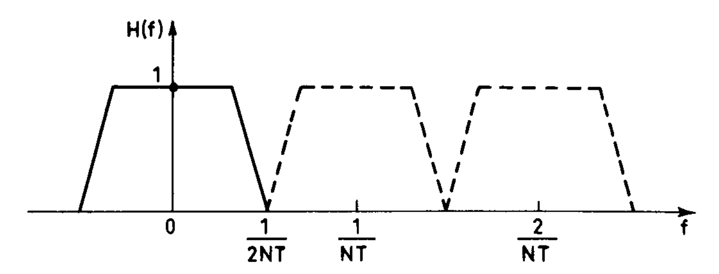


FIG. 10.16. *Fonction de filtrage d'un banc de filtres sans recouvrement*

La réponse impulsionnelle du filtre de base est la fenêtre d'analyse spectrale du système. Si les filtres sont spécifiés pour n'avoir aucun recouvrement en fréquence, la fréquence d'échantillonnage en sortie des filtres, ou en entrée suivant le mode d'utilisation, peut prendre la valeur $\frac{1}{NT}$ et l'ensemble des calculs peut être effectué à cette cadence.

Si N est une puissance de 2, l'algorithme de TFR peut être utilisé pour calculer la TFD et le nombre de multiplications réelles M_R à faire par période d'échantillonnage dans l'ensemble du système s'écrit :

$$M_R = N.2K + 2N\,\mathrm{Log}_2\,\frac{N}{2} = 2N\left[K + \mathrm{Log}_2\,\frac{N}{2}\right] \qquad (10.42)$$

Cette valeur est à comparer à la valeur $2K.N^2$ que nécessite un ensemble de N filtres RII de même ordre fonctionnant à la cadence $\frac{1}{T}$.

10.8 CONCLUSION

Les filtres RIF possèdent une propriété d'indépendance entre les cadences d'échantillonnage en entrée et en sortie, qui est exploitée pour adapter la cadence des calculs à la largeur de bande du signal au cours du traitement. Cette propriété s'étend aux structures récursives, moyennant une transformation. La fonction de base qui intervient est la fonction de déphaseur.

Le filtrage multicadence s'applique aux filtres à bande passante étroite. Il peut apporter des gains en calcul considérables quand existe entre la fréquence d'échantillonnage et la bande passante d'un filtre un rapport qui dépasse un ordre de grandeur, situation fréquente en pratique.

L'emploi de ces techniques impose une analyse plus fine du traitement. Les limitations à leur emploi proviennent principalement des complications dans l'enchaînement des calculs qu'apporte la mise en cascade d'étages différents et fonctionnant à des cadences différentes. Dans chaque application potentielle du filtrage

multicadence il faut examiner avec soin ce point pour éviter qu'un accroissement excessif de l'unité de commande ou du programme d'instructions ne vienne compenser les gains en calcul.

BIBLIOGRAPHIE

[1] M. Bellanger, J. Daguet and G. Lepagnol – Interpolation, Extrapolation and Reduction of Computation Speed in digital Filters. *IEEE Transactions,* ASSP-22, n° 4, Aug. 1974.

[2] F. Mintzer and B. Liu – Aliasing error in the design of multirate filters. *IEEE Transactions*, ASSP-26, Feb. 1978.

[3] D. J. Goodman and M. J. Carey – Nine Digital filters for decimation and Interpolation. *IEEE Transactions,* ASSP-25, April 1977.

[4] R. E. Crochière and L. R. Rabiner – Chapter 5 in : *Multirate Digital Signal Processing :* Prentice-Hall Inc., Englewood Cliffs, N. J., 1983.

[5] M. Bellanger, G. Bonnerot and M. Coudreuse – Digital Filtering by Polyphase Network : application to sample rate alteration and Filter Banks. *IEEE Transactions,* ASSP-24, n° 2, April 1976.

[6] P. P. Vaidyanathan – Multirate Systems and Filter Banks, Prentice Hall, 1993.

[7] N. J. Fliege – *Multirate Digital Signal Processing,* John Wiley, Chichester, 1994.

EXERCICES

1 Donner le nombre de bits affectés aux coefficients des filtres demi-bande du tableau 10.1.

Placer sur les abaques de la figure 10.5, les filtres F_6, F_8 et F_9.

Utiliser les résultats du paragraphe 5.10 pour évaluer le supplément d'ondulation apporté par la limitation du nombre de bits des coefficients dans les filtres demi-bande. Tester l'évaluation sur les filtres F_6, F_8 et F_9.

2 Estimer le nombre de coefficients des trois filtres de la cascade dans l'exemple 1 du paragraphe 10.4.

Chaque résultat de multiplication étant arrondi, analyser le bruit de calcul produit dans la réduction de fréquence d'échantillonnage et donner une expression de sa puissance. Même question pour l'élévation de fréquence d'échantillonnage et le filtre de base. Comparer avec la réalisation directe.

Donner une estimation de la distorsion de repliement.

3 Un filtre pour une voie téléphonique a une bande passante qui s'étend de 0 à 3400 Hz, avec une ondulation inférieure à 0,25 dB. L'affaiblissement est supérieur à 35 dB à partir de 4000 Hz. Pour une fréquence d'échantillonnage f_e = 32 kHz, donner une réalisation en filtrage multicadence avec une cascade de filtres demi-bande. Comparer le nombre de multi-

plications et d'additions à faire par seconde et le volume de mémoire avec les valeurs obtenues pour un filtrage RIF direct.

Le signal appliqué au filtre est composé de nombres à 13 bits, les calculs sont faits à l'aide de registres à 16 bits. Évaluer la puissance du bruit de calcul dans la réduction et dans l'élévation de fréquence d'échantillonnage.

4 Un calculateur de Transformée de Fourier Discrète est un banc de filtres uniforme dont les caractéristiques ont été indiquées au paragraphe 2.4. Étudier les déphasages introduits dans les Transformées de Fourier Impaire et doublement impaire du paragraphe 3.3. Comment sont caractérisés les bancs de filtres ainsi obtenus ?

5 Soit à réaliser un banc de deux filtres à partir du filtre RII d'ordre 4 donné en exemple au paragraphe 7.2.4. Les zéros et les pôles du demi-plan supérieur ont pour affixes (fig. 7.6)

$$Z_1 = -0{,}816 + j\,0{,}578\,; \quad Z_2 = -0{,}2987 + j\,0{,}954$$
$$P_1 = 0{,}407 + j\,0{,}313\,; \quad P_2 = 0{,}335 + j\,0{,}776$$

Utiliser les formules du paragraphe 10.6 pour calculer les fonctions de transfert en Z des branches du réseau polyphasé. Donner les affixes des pôles et zéros dans le plan complexe.

Utiliser les résultats du chapitre 7 pour déterminer l'incidence sur la réponse en fréquence de la limitation du nombre de bits des coefficients du dénominateur de la fonction de transfert. Comparer avec la réalisation directe.

Donner le schéma de réalisation du banc de deux filtres et compter le nombre de multiplications nécessaire en supposant que la fréquence d'échantillonnage en sortie de chaque filtre est la moitié de la valeur à l'entrée.

Chapitre 11

Filtres QMF et ondelettes

La compression de certains signaux, notamment la parole, le son et l'image, fait appel à la décomposition en sous-bandes avec réduction de fréquence d'échantillonnage et à la reconstitution à partir des sous-bandes, après stockage ou transmission. L'approche la plus simple pour réaliser ces opérations consiste à faire appel à des bancs de 2 filtres.

11.1 DÉCOMPOSITION EN DEUX SOUS-BANDES ET RECONSTITUTION

Le schéma d'ensemble est donné à la figure 11.1. Le signal $x(n)$ à analyser est appliqué à deux filtres, un passe-bas $H_0(z)$ et un passe-haut $H_1(z)$. Les sorties sont sous-échantillonnées par le facteur 2. La reconstitution est effectuée à partir de séquences dont un échantillon sur 2 est nul, filtrées l'une par un passe-bas $G_0(Z)$ et l'autre par un passe-haut $G_1(Z)$.

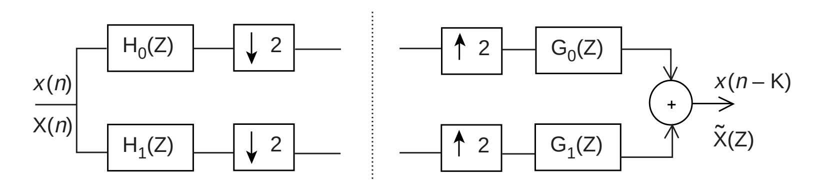


FIG. 11.1. *Bancs de 2 filtres pour décomposition-reconstitution par le facteur 2*

Comme expliqué au paragraphe 10.1, le sous-échantillonnage en sortie des filtres d'analyse provoque l'addition au signal utile des bandes images $H_0(-Z)X(-Z)$ et $H_1(-Z)X(-Z)$. La disparition de ces signaux indésirables en sortie du banc de filtres de synthèse implique la relation :

$$G_0(Z)H_0(-Z) + G_1(Z)H_1(-Z) = 0 \qquad (11.1)$$

Cette contrainte est satisfaite en utilisant les mêmes filtres dans les deux sous-ensembles et en prenant, pour les filtres de synthèse :

$$G_0(Z) = H_1(-Z)\,;\, G_1(Z) = -H_0(-Z) \qquad (11.2)$$

Alors, la condition de reconstitution s'exprime par :

$$H_0(Z)H_1(-Z) - H_1(Z)H_0(-Z) = Z^{-K} \qquad (11.3)$$

où K est le retard global apporté au signal.

Ensuite, il faut calculer les coefficients des filtres, en fonction du type de filtre et des spécifications dans le domaine des fréquences, le nombre de coefficients déterminant la qualité de la séparation entre les sous-bandes.

L'approche la plus simple, celle qui minimise la quantité de calculs, consiste à prendre des filtres à phases linéaires et identiques [1].

11.2 FILTRES QMF

Comme indiqué au chapitre précédent, un réseau polyphasé à 2 branches permet d'obtenir, avec les mêmes calculs, un filtre passe-bas et un filtre passe-haut. Le sous-échantillonnage peut alors être réalisé à l'entrée des filtres d'analyse et le schéma global est donné à la figure 11.2.

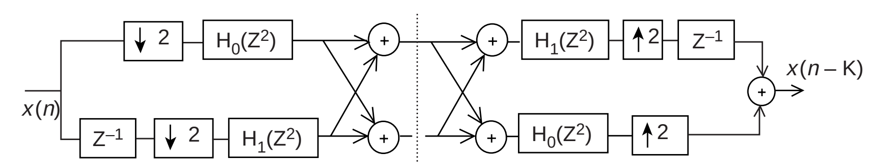


FIG. 11.2. *Filtre QMF*

Les fonctions de transfert $H_0(Z^2)$ et $H_1(Z^2)$ constituent la décomposition polyphase du filtre prototype H(Z) c'est-à-dire que :

$$H(Z) = H_0(Z^2) + Z^{-1}H_1(Z^2) \qquad (11.4)$$

Il faut ensuite déterminer les conditions que doit satisfaire le filtre prototype pour que les relations de base (11.1) et (11.2) soient vérifiées. La fonction de transfert totale du système a pour expression :

$$T(Z) = Z^{-1}H_1(Z^2)H_0(Z^2) \qquad (11.5)$$

Or, les composantes polyphases sont liées au filtre prototype par les expressions données au paragraphe 10.1, soit :

$$H_0(Z^2) = \frac{1}{2}\,[H(Z) + H(-Z)]\ ;\ Z^{-1}H_1(Z^2) = \frac{1}{2}\,[H(Z) - H(-Z)]$$

et, par suite :

$$T(Z) = \frac{1}{4}\,[H^2(Z) - H^2(-Z)] \tag{11.6}$$

On montre alors que, pour obtenir le changement de signe nécessaire à la reconstitution dans (11.2), il faut prendre un filtre prototype à nombre pair de coefficients : N = 2P. En effet, la réponse en fréquence d'un tel filtre s'écrit, comme indiqué au paragraphe 5.2 :

$$H(f) = e^{-j2\pi f\left(P - \frac{1}{2}\right)}\, 2\, H_R(f) \tag{11.7}$$

où $H_R(f)$ est une fonction réelle paire. On a également,

$$H\left(f - \frac{1}{2}\right) = e^{-j2\pi f\left(P - \frac{1}{2}\right)}\, 2\, H_R\left(f - \frac{1}{2}\right) \tag{11.8}$$

Dans ces conditions, sur le cercle unité, il vient :

$$\frac{1}{4}\,[H^2(Z) - H^2(-Z)]_{Z = e^{j2\pi f}} = e^{-j2\pi f\left(2P - 1\right)}\left[H_R^2(f) + H_R^2\left(\frac{1}{2} - f\right)\right] \tag{11.9}$$

et la condition de reconstitution s'écrit :

$$H_R^2(f) + H_R^2\left(\frac{1}{2} - f\right) = 1 \tag{11.10}$$

Le terme de phase dans l'expression (11.9) donne le retard apporté par la fonction de transfert totale et, donc, l'ensemble de décomposition-reconstitution, soit : K=2P−1.

La terminologie QMF (Quadrature Mirror image Filter) provient de la décomposition polyphase du filtre prototype à nombre pair de coefficients. En effet, comme expliqué au chapitre 5, ce filtre est interpolateur, c'est-à-dire qu'il apporte un retard d'un multiple impair de la demi-période d'échantillonnage. Dans ces conditions, les expressions obtenues au paragraphe 10.1 conduisent à la décomposition :

$$H(Z) = Z^{-\frac{1}{2}}\, H_{\frac{1}{2}}(Z^2) + Z^{-\frac{3}{2}}\, H_{\frac{3}{2}}(Z^2)$$

qui se réécrit plus simplement :

$$H(Z) = Z^{\frac{1}{2}}\,[H_0(Z^2) + Z^{-1} H_1(Z^2)] \tag{11.11}$$

Il vient alors pour les composantes polyphases :

$$H_0(Z^2) = Z^{\frac{1}{2}}\,[H(Z) + jH(-Z]\ ;\ H_1(Z^2) = Z^{\frac{1}{2}}\,[H(Z) - jH(-Z] \tag{11.12}$$

et la réponse en fréquence :

$$|H_0(f)| = \left|H(f) + jH\left(\frac{1}{2} - f\right)\right| \tag{11.13}$$

La bande de base et la bande image sont en quadrature. Par extension, le terme QMF peut s'appliquer à tous les bancs de 2 filtres pour décomposition-reconstitution.

Il reste, ensuite, à calculer les coefficients. La relation (11-10) montre que la fonction de transfert $H^2(Z)$ est celle d'un filtre demi-bande à nombre impair de coefficients de type Nyquist et $H(Z)$ est un filtre passe-bas demi-Nyquist. Le calcul se fait à partir d'un gabarit spécifiant pour $H(f)$ les limites de la bande passante f_1 et de la bande affaiblie f_2, les ondulations δ_1 et δ_2 et en imposant l'amplitude $\sqrt{2}/2$ à la fréquence 1/4. Pour que $H^2(f)$ approche la condition (11.10) avec une ondulation δ, on prend :

$$f_2 = \frac{1}{2} - f_1 \ ; \ H(0) = 1 \ ; \ \delta_1 = \frac{\delta}{2} \ ; \ \delta_2 = \sqrt{\delta}$$

Exemple :

Soit un filtre passe-bas à N = 8 coefficient dont les paramètres sont les suivants :

$$\Delta f = 0{,}24 \ ; \ f_1 = 0{,}13 \ ; \ f_2 = 0{,}37 \ ; \ \delta = 0{,}01$$

Le calcul fournit les coefficients suivants :

$$h_1 = h_8 = 0{,}015235 \ ; \ h_2 = h_7 = 0{,}085187$$
$$h_3 = h_6 = 0{,}081638 \ ; \ h_4 = h_5 = 0{,}486502$$

Les deux branches du réseau polyphasé obtenu, $H_0(Z^2)$ et $H_1(Z^2)$ ont les mêmes coefficients, mais dans l'ordre inverse.

Le filtre $\sum_{i=1}^{15} h'_i Z^{-i}$ devrait avoir ses coefficients pairs nuls, sauf h'_8.

En fait, on trouve $\sum_{i=1;\, i \neq 4}^{7} (h'_{2i})^2 = 1{,}7 \,.\, 10^{-5}$.

Des méthodes itératives peuvent, si nécessaire, compléter le calcul et permettre de mieux approcher la symétrie.

La qualité de la reconstitution dépend de l'ondulation δ et, donc, du nombre de coefficients. Pour obtenir une reconstitution parfaite, il faut abandonner l'option de filtres identiques et le sous-échantillonnage à l'entrée de l'analyse [2].

11.3 DÉCOMPOSITION ET RECONSTITUTION PARFAITE

En posant $P(Z) = H_0(Z)H_1(-Z)$, on observe que la condition de reconstitution (11.3) s'écrit :

$$P(Z) - P(-Z) = Z^{-K} \qquad (11.14)$$

C'est-à-dire que $P(Z)$ est un filtre passe-bas à nombre impair de coefficients dont tous les coefficients d'indice pair sont nuls, sauf le coefficient central qui est égal à l'unité. Par exemple, pour $M = 2$ coefficients différents, il vient :

$$P(Z) = h_3 + h_1 Z^{-2} + Z^{-3} + h_1 Z^{-4} + h_3 Z^{-6} \qquad (11.15)$$

et les coefficients des filtres passe-bas $H_0(Z)$ et $H_1(-Z)$ sont soumis à la contrainte que, dans leur produit, les coefficients des termes en Z^{-1} et Z^{-5} sont nuls. Il reste des degrés de liberté qui sont utilisés pour obtenir des propriétés particulières.

Dans un premier exemple, on choisit des polynômes de degré différents, à coefficients symétriques et ayant des zéros au point $Z=1$. Alors, en prenant :

$$H_1(-Z) = \frac{1}{2}(1+Z^{-1})^2$$
$$H_0(Z) = (1+Z^{-1})(\alpha_0 + \alpha_1 Z^{-1} + \alpha_1 Z^{-2} + \alpha_0 Z^{-3}) \qquad (11.16)$$

la condition de reconstitution parfaite impose que, dans le produit $P(Z)$, le coefficient du terme Z^{-1} soit nul et celui du terme Z^{-3} égal à l'unité, ce qui donne $\alpha_0 = -\frac{1}{8}$ et $\alpha_1 = \frac{3}{8}$.

Finalement :

$$H_0(Z) = \frac{1}{8}[-1 + 2Z^{-1} + 6Z^{-2} + 2Z^{-3} - Z^{-4}] \qquad (11.17)$$

Les deux filtres ainsi obtenus sont à la base de la transformation réversible utilisée dans la norme de compression des images fixes JPEG 2000, dans l'option sans perte. Les réponses en fréquence des filtres sont données à la figure 11.3. Il faut noter le déséquilibre des deux sous-bandes, les filtres ne sont pas du type demi-Nyquist.

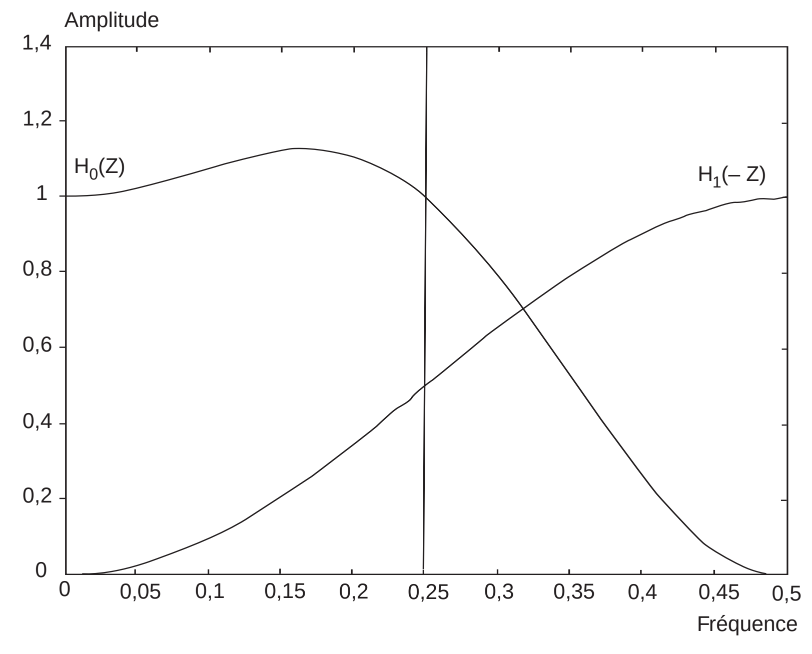


FIG. 11.3. *Réponse en fréquence des filtres dans la norme JPEG sans perte*

Une décomposition en deux sous-bandes égales peut être obtenue en abandonnant l'option de phase linéaire et en factorisant comme suit le filtre demi-bande :

$$P(Z) = H_0(Z)\,Z^{-k}\,H_0(Z^{-1}) \qquad (11.18)$$

c'est-à-dire que les filtres $H_0(Z)$ et $H_1(-Z)$ ont les mêmes coefficients mais dans l'ordre inverse et le polynôme $P(Z)$ est de degré 2K et possède 2K +1 coefficient. En fonction du nombre M de coefficients différents dans le filtre demi-bande, on a l'égalité $2K+1=4M$, c'est-à-dire $K = 2M - 1$. Le fait que l'entier K soit impair permet de satisfaire la relation (11.2) et, avec $G_0(Z) = Z^{-k}H_0(Z^{-1})$, $H_1(Z) = -Z^{-k}H_0(Z^{-1})$ et $G_1(Z) = -H_0(-Z)$, on vérifie que les conditions (11.1) et (11.3) sont satisfaites.

La procédure de calcul des coefficients est celle des filtres RIF à phase minimale décrite au paragraphe 5.13. Au coefficient central d'un filtre demi-bande, on ajoute l'ondulation, ce qui rend les zéros sur le cercle unité doubles, puis, les facteurs à phases minimales et maximales sont extraits [3].

Comme exemple, on considère le cas d'un filtre P(Z) à 2K + 1 = 15 coefficients, calculé avec les spécifications $f_1 = \frac{1}{2} - f_2 = 0{,}2$. Les M = 4 coefficients différents ont pour valeurs :

$$h_1 = 0{,}62785 \; ; \; h_3 = 0{,}18681 \; ; \; h_5 = 0{,}08822 \; ; \; h_7 = 0{,}05297$$

L'ondulation a pour valeur $\delta = 0{,}047$ et le coefficient central devient : $h_0 = 1{,}047$.

En prenant un des zéros qui sont sur le cercle unité et ceux qui sont à l'intérieur, on obtient pour le premier filtre :

$$H_0(Z) = 0{,}3704 + 0{,}5111\,Z^{-1} + 0{,}2715\,Z^{-2} - 0{,}0885\,Z^{-3} - 0{,}1346\,Z^{-4} + 0{,}0338\,Z^{-5} + 0{,}0973\,Z^{-6} - 0{,}0703\,Z^{-7}$$

La réponse en fréquence correspondante est donnée à la figure 11.4. Par rapport au filtre d'origine, l'ondulation en bande affaiblie est devenue $\sqrt{\delta}$.

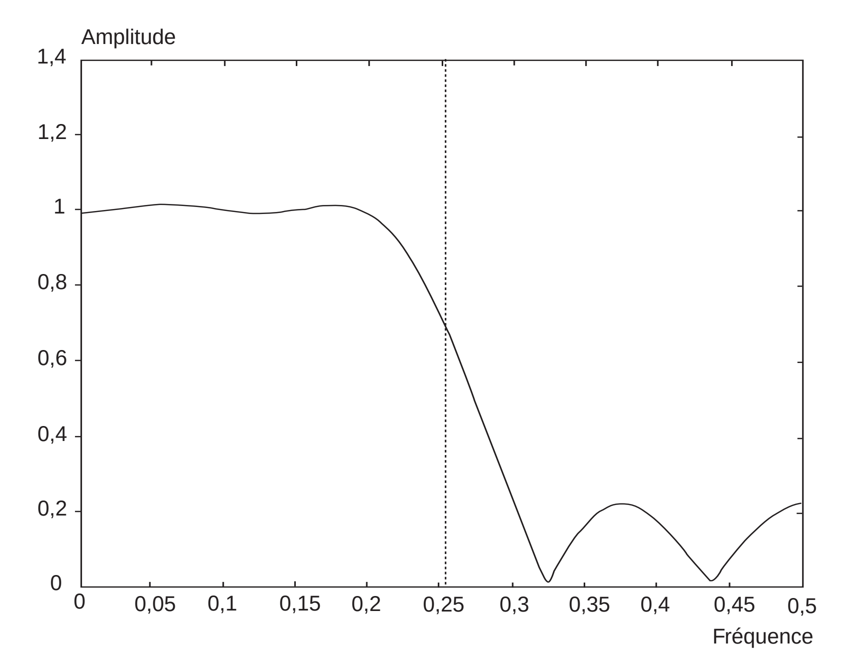


FIG. 11.4. *Réponse en fréquence du filtre d'analyse passe-bas*

Dans les caractéristiques à retenir pour les facteurs de $P(Z)$, il en est une qui est intéressante pour la compression des signaux et des images en particulier, c'est la régularité de la réponse en fréquence. En filtrage, cette caractéristique se traduit par la présence de zéros multiples au point $Z = -1$ dans la fonction de transfert. Sur le plan des principes, l'approche est justifiée par la théorie des ondelettes.

11.4 ONDELETTES

La théorie des ondelettes a pour objectif la représentation des signaux dans le domaine temps-fréquence. C'est une représentation que l'analyse de Fourier ne permet pas, car elle suppose le signal périodique ou de durée infinie. Ainsi, pour localiser un signal à la fois dans le temps et en fréquence avec la transformée de Fourier, il faut introduire une fenêtre glissante.

La transformée en ondelette utilise comme base de décomposition des fonctions dites ondelettes, déduites d'une fonction génératrice par translation et dilatation. Elle permet d'analyser des signaux de durée quelconque [4, 5].

En pratique, la transformée en ondelettes discrète correspond à un banc de filtres non uniforme et une approche efficace pour la mise en œuvre consiste à mettre en cascade des bancs de 2 filtres comme ceux des paragraphes précédents, avec réduction par 2 des fréquences d'échantillonnage à chaque étage et mêmes coefficients pour les filtres. Les opérations de translation et dilatation sont effec-

tuées automatiquement par les changements de fréquence d'échantillonnage. La structure en arbre ainsi obtenue est illustrée à la figure 11.5 pour l'analyse. Bien entendu, on peut aussi obtenir un banc de filtres uniforme en complétant la branche basse sur le schéma.

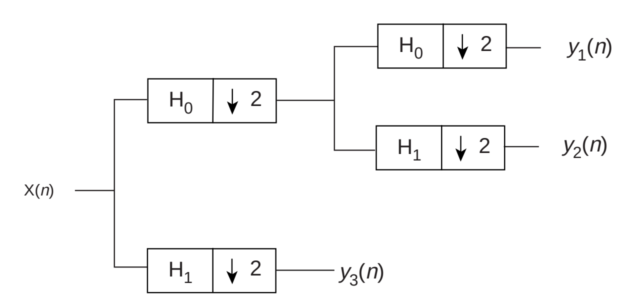


FIG. 11.5. *Banc de filtres non uniforme, par structure en arbre*

Pour le calcul des coefficients, l'objectif est d'atteindre la régularité maximale, c'est-à-dire d'avoir dans la fonction P(Z) le maximum de zéros au point Z = – 1. Un filtre demi-bande à N coefficients différents possède 4N – 1 coefficients en tout et est de degré 4N – 2. Ce filtre ayant 2(N – 1) coefficients nuls, la fonction P(Z) comporte un facteur de degré 2(N – 1). Alors, le facteur restant est au maximum de degré 2N. Ensuite, il faut factoriser P(Z) pour obtenir les deux filtres d'analyse et synthèse.

Dans une solution à phase minimale, les filtres obtenus possèdent 2N coefficients et la fonction de transfert en Z du filtre passe-bas présente N zéros au point Z = – 1 du plan complexe. Comme dans les paragraphes précédents, les valeurs numériques peuvent être déterminées directement en combinant cette contrainte avec les conditions d'annulation des coefficients des termes impairs dans le produit P(Z). On peut aussi obtenir P(Z) et factoriser.

Le tableau 11.1 donne les coefficients des filtres $H_0(Z)$ pour les premières valeurs de N. Les coefficients des autres filtres intervenant dans l'analyse et la synthèse, $H_1(Z)$, $G_0(Z)$ et $G_1(Z)$ selon la figure 11.1, sont donnés par :

$$H_0(Z) = \sum_{i=1}^{2N} h_{0,i} Z^{-i} \ ; h_{1,i} = (-1)^i h_{0,2N+1-i} \ ; \ g_{1,i} = h_{1,2N+1-i} \qquad (11.19)$$

Les réponses en fréquence sont données à la figure 11.6. On remarque la similitude avec les réponses des filtres de Butterworth du paragraphe 7.2.3, qui possèdent les mêmes zéros au point Z = –1, ont également la propriété de reconstitution parfaite quand la fréquence de coupure est placée au milieu de la bande utile et ont une réponse en fréquence monotone.

Tableau 11.1. – Coefficients des ondelettes à phase minimale

N = 2	N = 3	N = 4	N = 5
0,482963	0,332671	0,230378	0,160102
0,836516	0,806892	0,714847	0,603828
0,224144	0,459878	0,630881	0,724307
– 0,129410	– 0,135011	– 0,027984	0,138427
	– 0,085441	– 0,187035	– 0,242295
	0,035226	0,030841	– 0,032245
		0,032883	0,077571
		– 0,010597	– 0,006242
			– 0,012581
			– 0,003336

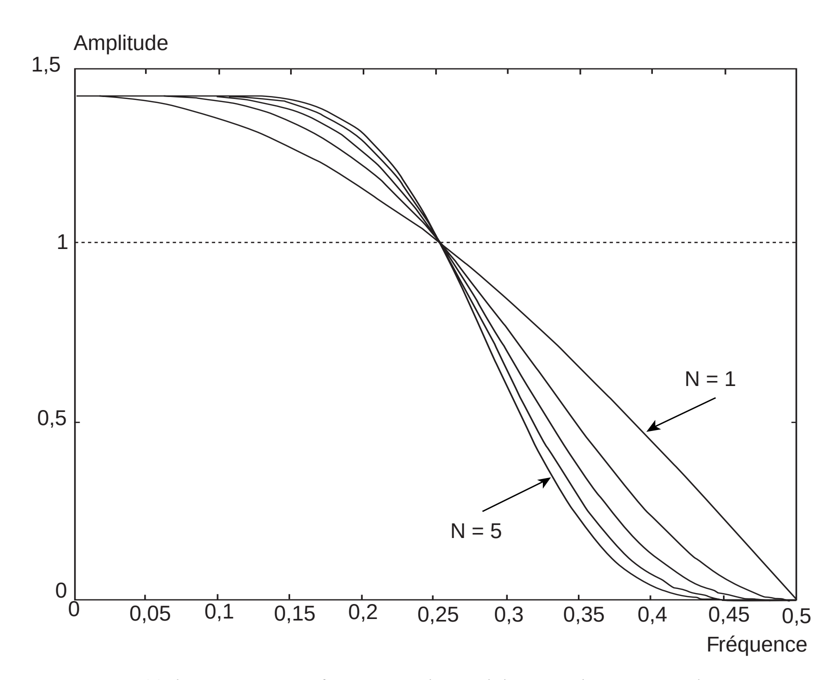


Fig. 11.6. *Réponse en fréquences des ondelettes à phase minimale*

Il est également possible d'obtenir des filtres à phase linéaire, en donnant aux filtres passe-bas et passe-haut des nombres de coefficients différents et impairs.

Par exemple, le tableau 11.2 donne les coefficients des filtres (9,7) utilisés dans la norme JPEG 2000, pour la compression avec taux élevé et pertes [6].

Tableau 11.2. – Filtres pour compression avec perte dans JPEG 2000

Z^i	$H_0(Z)$	$H_1(Z)$
i = ± 0	0,602949	1,115087
i = ± 1	0,266864	– 0,591272
i = ± 2	– 0,078223	– 0,057544
i = ± 4	– 0,016864	0,091272

L'évaluation de la précision de reconstitution se fait par le calcul de la réponse impulsionnelle de l'ensemble analyse-synthèse. En multipliant par $H_1(-Z)$ le polynôme obtenu en annulant les coefficients d'indice impair dans $H_0(Z)$ et en ajoutant le produit par $H_0(-Z)$ du polynôme obtenu en annulant les coefficients d'indice pair dans $H_1(Z)$, on vérifie que le polynôme obtenu a son coefficient central égal à l'unité et ses autres coefficients nuls, avec un écart inférieur à 2.10^{-6}. Cet écart provient de l'arrondi des valeurs des coefficients.

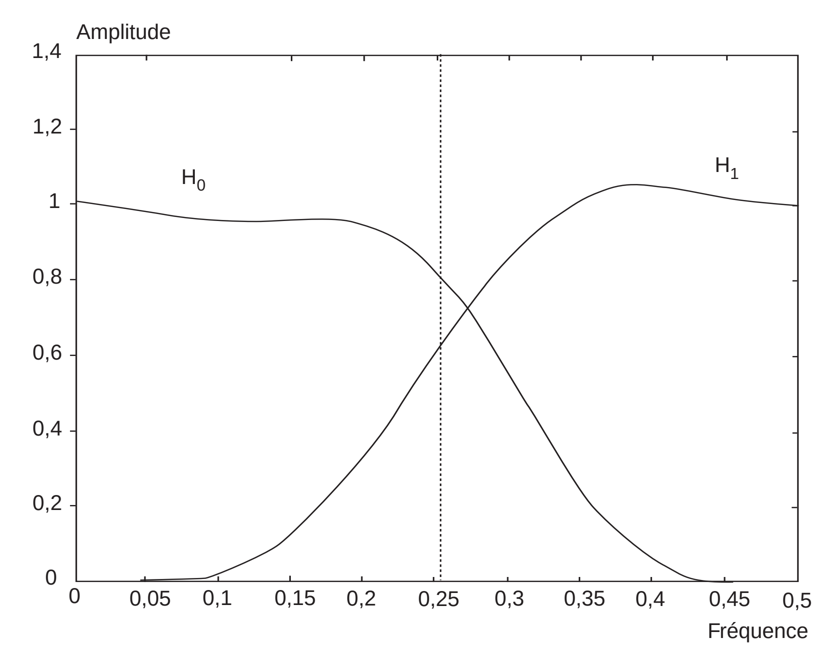


Fig. 11.7. *Réponse en fréquence des filtres dans la norme JPEG 2000 avec perte*

La figure 11.7 montre les réponses en fréquence des filtres $H_0(Z)$ et $H_1(Z)/2$, qui font encore apparaître un déséquilibre dans le partage du signal, mais plus faible que celui de la figure 11.3, les filtres ayant davantage de coefficients. Ces filtres ont la moitié de leurs zéros aux points $Z = \pm 1$ du plan complexe, ce qui donne une grande régularité aux réponses en fréquence, propriété importante en traitement d'images. Cependant, il faut signaler que cette approche produit une

amplification du bruit à la synthèse. En effet, dans la compression des images, une opération de quantification a lieu après l'analyse et la synthèse apporte une amplification du bruit de quantification par le facteur 1,243. Ce facteur est plus élevé que pour d'autres techniques de calcul des filtres QMF.

En ce qui concerne la complexité arithmétique, il faut souligner que le nombre de multiplications est très proche de celui de la technique polyphase, car il est possible de profiter de la symétrie des coefficients, à la fois dans la partie analyse et dans la partie synthèse.

11.5 STRUCTURE EN TREILLIS

La factorisation (11.8) du filtre demi-bande peut également se faire avec la représentation en treillis, en annulant un coefficient sur deux. La structure modulaire obtenue est donnée à la figure 11.8.

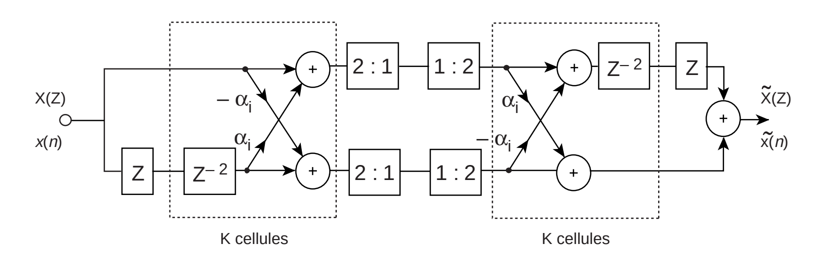


Fig. 11.8. *Banc de 2 filtres en structure treillis*

Les fonctions de transfert des filtres d'analyse et synthèse sont obtenues par les calculs indiqués au paragraphe 8.5. Par exemple, pour 3 cellules, K = 3, il vient :

$$H_0(Z) = 1 + \alpha_1 Z^{-1} - \alpha_1\alpha_2 Z^{-2} + \alpha_2 Z^{-3}$$
$$H_1(Z) = -\alpha_2 - \alpha_1\alpha_2 Z^{-1} - \alpha_1 Z^{-2} + Z^{-3} \quad (11.20)$$

ce qui correspond à la relation donnée précédemment, soit $H_1(Z) = -Z^{-K} H_0(-Z^{-1})$. La fonction de transfert de l'ensemble s'écrit :

$$T(Z) = Z^{-2(K-1)} \prod_{i=1}^{K} (1 + \alpha_i^2) \quad (11.21)$$

Le retard est donc de $2(K-1)$ périodes d'échantillonnage. Les coefficients du treillis sont calculés à partir des spécifications sur les filtres. Par exemple, en imposant un zéro au point $Z = -1$ et $H_0(1) = 4$, on obtient : $\alpha_1 = 1 + \sqrt{2}$; $\alpha_2 = 1 - \sqrt{2}$. En comparant avec le paragraphe précédent, il apparaît que le treillis est moins efficace que les autres factorisations, puisqu'au tableau 11.1 le filtre à 4 coefficients possède 2 zéros au point $Z = -1$. Cependant, il présente des avantages en

réalisation, notamment la modularité de la structure et le fait que la propriété de reconstitution parfaite n'est pas affectée par l'arrondi des coefficients [7].

BIBLIOGRAPHIE

[1] R.E.Crochière and L.R.Rabiner, « Multirate Digital Signal Processing », Prentice-Hall Inc., Englewood Cliffs, N.J., 1983.

[2] M. Vetterli, « Filter Banks Allowing Perfect Reconstruction », Signal Processing, Vol. 10, n° 3, April 1986, pp. 219-244.

[3] M. Smith and T. Barnwell, « Exact Reconstruction Techniques for Tree Structured Sub-band Coders », IEEE Transactions, ASSP-34, n° 3, June 1986, pp. 434-441.

[4] I. Daubechies, « Orthonormal Bases for Compactly Supported Wavelets », Communications on Pure and Applied Mathematics, Vol. 41, 1988, pp. 909-996.

[5] S. Mallat, « A Wavelet Tour of Signal Processing », 2nd Ed., New York: Academic, 1999.

[6] A. Skodras, C. Christopoulos and T. Ebrahimi, « The JPEG 2000 Still Image Compression Standard », IEEE Signal Processing Magazine, Vol. 18, n° 5, Sept. 2001, pp. 36-58.

[7] P.P. Vaidyanathan, « Multirate Systems and Filter Banks », Prentice Hall Inc. Englewood Cliffs, N.J. 1993.

EXERCICES

1 On applique le signal $x(n) = \cos(n\pi/4)$ à un filtre de fonction de transfert :

$$H(Z) = -0{,}050 + 0{,}117\, Z^{-1} + 0{,}452\, Z^{-2} + 0{,}452\, Z^{-3} + 0{,}117\, Z^{-4} - 0{,}050\, Z^{-5}$$

Quel est le retard apporté par ce filtre et quelle est sa réponse en fréquence. Donner l'expression du signal de sortie.

Donner le schéma d'un banc de 2 filtres QMF ayant H (Z) comme filtre prototype. Exprimer les signaux en sortie des filtres d'analyse. Calculer la fonction de transfert de l'ensemble analyse-synthèse et donner l'expression du signal reconstitué.

2 Dans une décomposition-reconstitution parfaite, on cherche à factoriser le polynôme prototype P (Z), supposé de degré 6, en deux facteurs de même degré. Recherchant la régularité maximale, on impose au filtre passe-bas $H_0(Z)$ d'avoir ses 3 zéros au point $Z = -1$. Calculer les coefficients du filtre passe-haut $H_1(Z)$ et donner les valeurs de ses zéros.

Comparer les réponses en fréquence avec celles de la figure 11.3.

Une quantification intervient en sortie des filtres d'analyse. Calculer le facteur d'amplification dans la phase de synthèse.

3 Des filtres d'analyse à phase linéaire possèdent 3 et 5 coefficients. Calculer les valeurs pour avoir la meilleure séparation possible des sous-bandes et la reconstitution parfaite.

Quand un bruit blanc de puissance unité est appliqué à l'entrée des filtres d'analyse, quelles puissances trouve-t-on en sortie.

4 Le signal $x(n) = \cos(n\pi/4)$ est appliqué aux filtres d'analyse du tableau 11.2. Donner l'expression des signaux en sortie des filtres, avant et après sous-échantillonnage. Dans la reconstitution, donner l'expression des signaux en sortie de chacun des filtres, avant l'addition finale.

5 On se propose d'étudier la sensibilité de l'opération de décomposition-reconstitution à la précision des coefficients.

Donner une borne supérieure de l'erreur de reconstitution due à l'arrondi des coefficients pour les filtres du tableau 11.1. Vérifier le résultat en prenant le premier filtre du tableau 11.1, à 4 coefficients, avec une représentation sur 4 bits. Donner la réponse en fréquence du filtre avec cette représentation.

Mêmes questions que ci-dessus pour les filtres du tableau 11.2.

Chapitre **12**

Bancs de filtres

Les techniques de décomposition et reconstitution des signaux présentées au chapitre précédent se généralisent à un nombre quelconque de sous-bandes avec l'utilisation de bancs de plus de deux filtres. En principe, dans ce cas également, le sous-échantillonnage peut intervenir en sortie des filtres d'analyse, mais on préfère généralement effectuer le sous-échantillonnage en entrée pour exploiter la combinaison réseau polyphasé-TFD et minimiser ainsi la complexité arithmétique.

12.1 DÉCOMPOSITION ET RECONSTITUTION

Dans la réalisation d'un banc de filtres par réseau polyphasé et TFD exposée au paragraphe (10.8), les opérations qui interviennent sont réversibles, ce qui conduit au schéma de la figure 12.1, pour une opération de décomposition et reconstitution d'un signal [1-2].

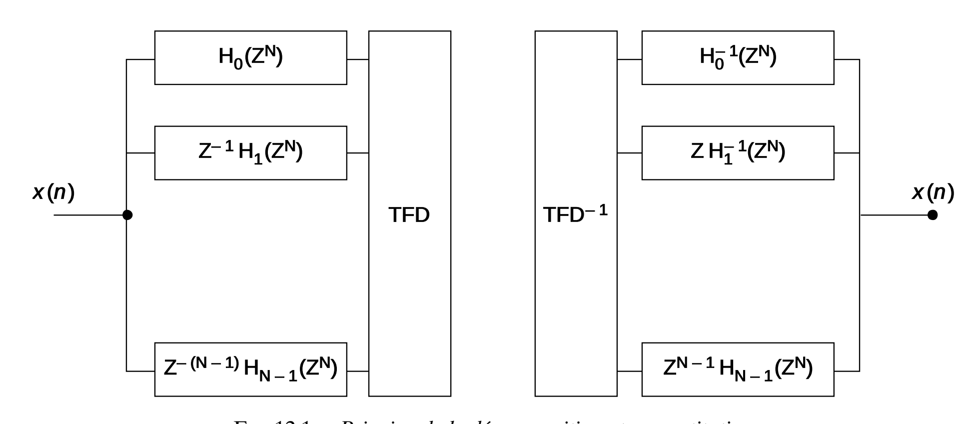


FIG. 12.1. *Principe de la décomposition et reconstitution*

La difficulté, dans la réalité, consiste à mettre en œuvre les opérations associées aux fonctions $H_i^{-1}(Z^N)$.

Le filtre H(Z) qui sert de base au processus, le filtre prototype, possède une décomposition polyphase dont les éléments satisfont la relation (10.8) :

$$Z^{-i}H_i(Z^N) = \frac{1}{N}\sum_{m=0}^{N-1} e^{-j\frac{2\pi}{N}im}\, H\left(Ze^{-j\frac{2\pi}{N}m}\right); \quad 0 \leqslant i \leqslant N-1 \quad (12.1)$$

Si le filtre de base H(Z) a une fréquence de coupure inférieure à $\frac{1}{2N}$ et s'il présente un affaiblissement important pour les fréquences supérieures ou égales à $\frac{1}{2N}$, c'est-à-dire si les repliements de spectre dûs à l'échantillonnage à la cadence $\frac{1}{N}$ sont négligeables, on peut écrire :

$$H\left[e^{j2\pi\left(f-\frac{m}{N}\right)}\right].H\left[e^{j2\pi\left(f-\frac{k}{N}\right)}\right] = 0; \qquad m \neq k \qquad (12.2)$$

Dans ces conditions l'égalité suivante est vérifiée sur le cercle unité, au facteur Z^{-N} près :

$$H_i(Z^N).H_{N-i}(Z^N) = \frac{1}{N^2}\sum_{m=0}^{N-1} H^2\left(Ze^{-j\frac{2\pi}{N}m}\right) = H_0^2(Z^N)$$

Soit :

$$\frac{H_i(Z^N)}{H_0(Z^N)}\cdot\frac{H_{N-i}(Z^N)}{H_0(Z^N)} = 1; \qquad 0 \leqslant i \leqslant N-1 \qquad (12.3)$$

Ces égalités traduisent simplement les relations de phase illustrées par la figure 10.15.

Alors, on peut prendre $H_{N-i}(Z^N)$ pour réaliser $H_i^{-1}(Z^N)$, c'est-à-dire utiliser le même banc de filtres à la décomposition et à la reconstitution du signal, l'opération globale correspondant à une multiplication par $H_0^2(Z^N)$.

Dans un certain nombre d'applications, il n'est pas possible de négliger les repliements de spectre; c'est le cas par exemple quand il faut décomposer un signal, le sous-échantillonner par le facteur N et le reconstituer ensuite avec la plus grande précision possible sur l'ensemble de la bande. Alors, soit G(Z) la fonction de transfert du filtre de base pour la reconstitution. Comme le produit d'une transformée de Fourier discrète par son inverse est égal à l'unité, l'opération globale correspond à une décomposition du signal $x(n)$ en N suites entrelacées $x(pN+i)$, auxquelles sont appliqués N opérateurs de fonction de transfert $G_i(Z^N).H_i(Z^N)$.

Avec la réduction de fréquence d'échantillonnage par N dans la partie décomposition, appelée aussi analyse, et l'augmentation par N dans la partie reconstitution, appelée aussi synthèse, le schéma correspondant est donné à la figure 12.2. Tous les traitements dans le dispositif correspondant se font à la cadence $\frac{1}{N}$, ce qui en fait une approche particulièrement efficace.

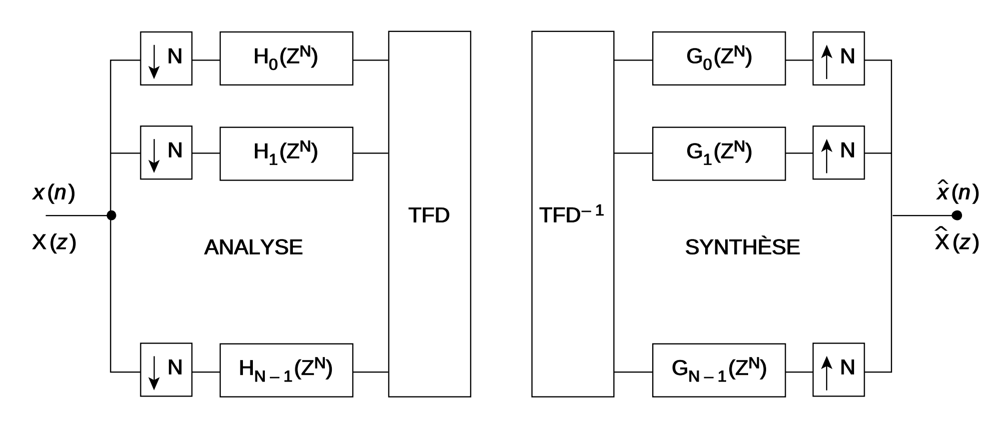


Fig. 12.2. *Bancs de filtres polyphasés pour l'analyse et la synthèse des signaux*

La condition de reconstitution avec le retard K s'écrit :

$$G_i(Z^N)\, H_i(Z^N) = Z^{-K}; \quad 0 \leqslant i \leqslant N-1 \tag{12.4}$$

Le retard K doit être le même dans toutes les branches du réseau polyphasé pour que l'entrelacement correspondant à l'augmentation de fréquence d'échantillonnage par N puisse rendre le signal initial retardé, $x(n-K)$.

Pour déterminer les fonctions inverses $G_i(Z^N)$, il faut procéder à une analyse détaillée de la réponse en fréquence des éléments du réseau polyphasé.

12.2 ANALYSE DES ÉLÉMENTS DU RÉSEAU POLYPHASÉ

La réponse en fréquence des éléments $H_i(Z^N)$ du réseau polyphasé se déduit directement de la relation (12.1). Cependant, des simplifications peuvent intervenir. En effet, les filtres du banc présentent généralement un recouvrement limité, par exemple, comme sur la figure 12.3. Dans ces conditions, un filtre donné ne se superpose qu'avec ses voisins immédiats si la réponse du filtre prototype H(Z) est telle que $H(f) = 0$ pour $|f| > \frac{1}{N}$. Alors, on peut écrire pour la branche d'indice i :

$$H_i(f) = H(f) + e^{-j\frac{2\pi}{N}i}\, H\left(f - \frac{1}{N}\right) \tag{12.5}$$

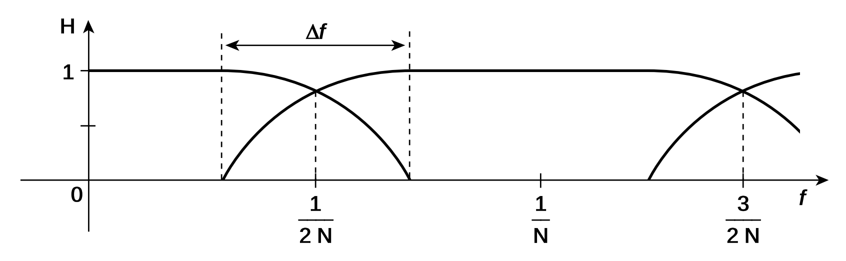


Fig. 12.3. *Recouvrement des filtres voisins*

La périodicité de cette réponse étant $\frac{1}{N}$, si les coefficients sont réels, il suffit de considérer la réponse dans l'intervalle $0 \leqslant f \leqslant \frac{1}{2N}$ et il vient :

$$H_i(f) = H(f) + e^{-j\frac{2\pi}{N}i}\,\overline{H}\left(\frac{1}{N} - f\right) \tag{12.6}$$

En supposant que la réponse du filtre prototype soit une courbe monotone décroissante dans la bande de transition Δf, $H_i(f)$ ne peut s'annuler pour $f \neq \frac{1}{2N}$.

Pour la valeur $\frac{1}{2N}$, il vient :

$$H_i\left(\frac{1}{2N}\right) = H\left(\frac{1}{2N}\right)\left(1 + e^{-j\frac{2\pi}{N}i}\right) \tag{12.7}$$

Cette réponse est nulle pour la branche $i = \frac{N}{2}$.

Donc, avec la décomposition (12.1), elle-même issue de l'expression (10.36), la branche d'indice $\frac{N}{2}$ n'est pas inversible, puisque sa fonction de transfert en Z possède un zéro dans le plan Z au point – 1. Pour obtenir un ensemble de branches inversibles, il faut faire appel à une autre décomposition polyphase.

Au chapitre V, il a été montré qu'un filtre RIF à phase linéaire à nombre de coefficients pair est un interpolateur à mi-période d'échantillonnage. On peut considérer qu'il provient d'un filtre RIF à nombre de coefficients impair, ayant la même réponse en fréquence, par un sous-échantillonnage d'un facteur 2. Il faut donc partir d'une décomposition polyphase à 2N branches et ne conserver qu'une branche sur deux. On obtient alors pour H (Z) la formule :

$$H(Z) = \sum_{i=0}^{N-1} Z^{-\left(i+\frac{1}{2}\right)} H_{i+\frac{1}{2}}(Z^N) \tag{12.8}$$

Avec cette décomposition, l'amplitude minimale H_{min} dans une branche est donnée par :

$$H_{min} = \left| H_{\frac{N}{2}+\frac{1}{2}}\left(\frac{1}{2N}\right) \right| = \left| \left(1 - e^{-j\frac{\pi}{N}}\right) \right| = 2 \sin \frac{\pi}{2N} \qquad (12.9)$$

Par suite, pour avoir un réseau polyphasé inversible, il suffit d'imposer au filtre prototype RIF à phase linéaire d'avoir un nombre pair de coefficients.

Quant à la position des zéros des fonctions $H_{i+\frac{1}{2}}(Z^N)$dans le plan Z^N par rapport au cercle unité, on observe qu'ils se répartissent également entre l'intérieur et l'extérieur du cercle unité. La justification se trouve dans le fait que les éléments $H_{i+\frac{1}{2}}(Z^N)$ sont à phase presque linéaire. De plus, l'amplitude de la réponse en fréquence restant proche de l'unité, les zéros sont loin du cercle unité, sauf pour les branches qui présentent un affaiblissement important à la fréquence $\frac{1}{2N}$ quand N est grand.

Comme exemple, on peut prendre le cas d'un banc de N = 2 branches avec un filtre prototype à 16 coefficients, selon le tableau 12.1.

Tableau 12.1. – COEFFICIENTS D'UN FILTRE PROTOTYPE

	h_i	coefficients
0	$h_0 = h_{15}$	0,002898
1	$h_1 = h_{14}$	– 0,009972
2	$h_2 = h_{13}$	– 0,001921
3	$h_3 = h_{12}$	0,035969
4	$h_4 = h_{11}$	– 0,016119
5	$h_5 = h_{10}$	– 0,095302
6	$h_6 = h_9$	0,106799
7	$h_7 = h_8$	0,477347

La fonction de transfert définie par :

$$H_0(Z^2) = \sum_{i=0}^{7} h_{2i}(Z^{-2})^i \qquad (12.10)$$

possède 7 zéros répartis comme indiqué sur la fig 12.4.

A l'évidence, ils sont éloignés du cercle unité. La seconde fonction, $H_1(Z^2)$, possède les mêmes coefficients, mais dans l'ordre inverse. Les zéros sont donc les inverses des précédents.

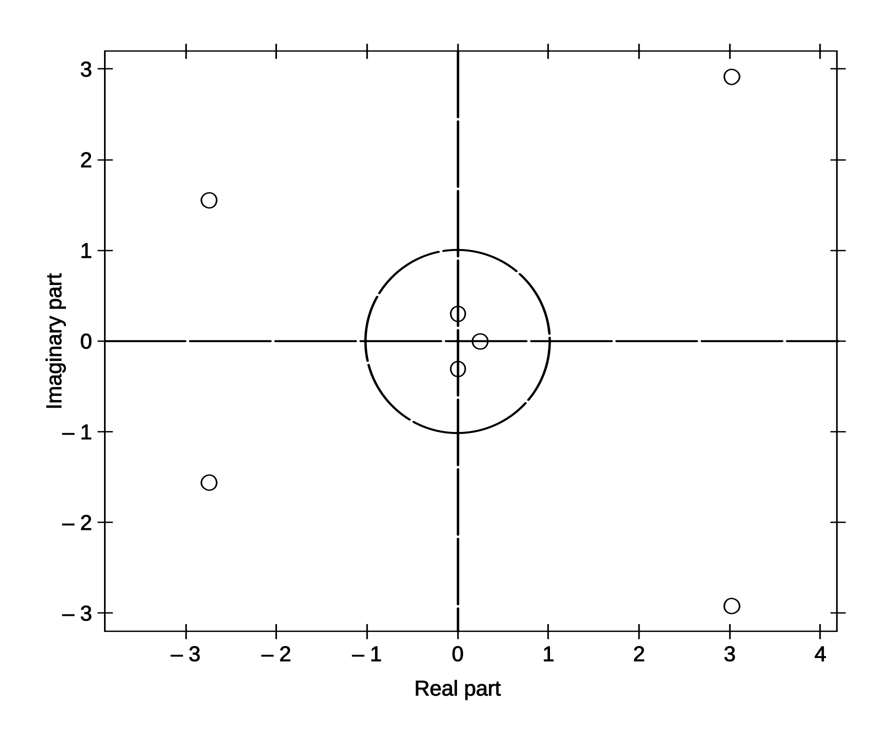


FIG. 12.4. *Zéros d'un élément polyphase avec* N = 2

12.3 CALCUL DES FONCTIONS INVERSES

En se plaçant dans le plan Z pour les fonctions de transfert, le calcul de la fonction inverse pour les éléments polyphases commence par une factorisation où les L_1 zéros à l'intérieur du cercle unité sont séparés des L_2 zéros qui sont à l'extérieur :

$$H_i(Z) = h_{i0} \prod_{k=1}^{L_1} (1 - Z_k Z^{-1}) \prod_{\ell=0}^{L_2} (1 - Z_\ell Z^{-1}) \qquad (12.11)$$

Le terme h_{i0} est un facteur d'échelle.

Pour tout zéro Z_ℓ extérieur au cercle unité, on peut écrire :

$$\frac{1}{1 - Z_\ell Z^{-1}} = \frac{-Z}{Z_\ell} \sum_{i=0}^{\infty} (Z_\ell^{-1})^i Z^i \qquad (12.12)$$

et, par suite, l'inverse du second facteur de (12.12) peut être approché avec une précision arbitraire par un nombre fini de termes. Soit la fonction $G_i(Z)$ définie par :

$$G_i(Z) = \frac{\sum_{l=0}^{L_3} a_l Z^{-l}}{\prod_{k=1}^{L_1} (1 - Z_k Z^{-1})} = \frac{\sum_{l=0}^{L_3} a_l Z^{-l}}{1 + \sum_{l=0}^{L_1} b_l Z^{-l}} \qquad (12.13)$$

où L_3 est un entier.

La condition d'inversion est satisfaite si :

$$\left(\sum_{l=0}^{L_2} c_l Z^{-l}\right)\left(\sum_{l=0}^{L_3} a_l Z^{-l}\right) = Z^{-(L_2 + L_3)} \qquad (12\text{-}14)$$

Le choix du retard $L_2 + L_3$ se justifie par le fait que les coefficients du développement (12-12) sont décroissants et que le second facteur dans (12.11) est à phase maximale. La relation d'inversion s'écrit sous forme matricielle :

$$MA = \begin{bmatrix} C_0 & 0 & 0 & \dots & 0 \\ C_1 & C_0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & C_{L_2} \end{bmatrix} \begin{bmatrix} a_0 \\ a_l \\ \vdots \\ a_{L_3} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \qquad (12.15)$$

où le vecteur A a pour éléments les coefficients a_l inconnus. Le système est surdéterminé et il admet une solution au sens des moindres carrés, donnée par la relation :

$$A = (M^t M)^{-1} M^t \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \qquad (12.16)$$

En reprenant l'exemple précédent, les coefficients a_l et b_l de $G_0(Z)$ qui interviennent dans l'équation (12.13) sont donnés dans le tableau 12.2, pour $L_3 = 5$.

Tableau 12.2. – Coefficients d'un élément polyphase de synthèse

	a_i	b_i
0	– 0,000017	1
1	– 0,001214	– 0,207030
2	0,000053	– 0,033224
3	0,000662	0,002642
4	– 0,004129	0,005796
5	– 0,019944	

L'écart par rapport à la valeur idéale de l'inverse, c'est-à-dire l'erreur de reconstitution se calcule comme la norme du vecteur $(MA)^t - [0, 0, \dots, 1]$. En valeur quadratique on obtient :

$$J = \|(MA)^t - [0, 0, \dots, 1]\|^2 \qquad (12.17)$$

Dans l'exemple ci-dessus, on trouve $J = 1{,}02.10^{-7}$ avec $L_3 = 5$ et $J = 8.10^{-7}$ avec $L_3 = 4$.

La structure des éléments polyphase de synthèse $G_i(Z)$ est celle d'un filtre RII général. Comme les pôles Z_k sont éloignés du cercle unité, une réalisation en structure directe est possible. Elle est représentée à la figure 12.5. Avec $L_3 = 5$, il faut donc 9 multiplications et 5 mémoires pour réaliser un élément et le retard apporté par l'ensemble analyse et synthèse est égal à $L_2 + L_3 = 9$.

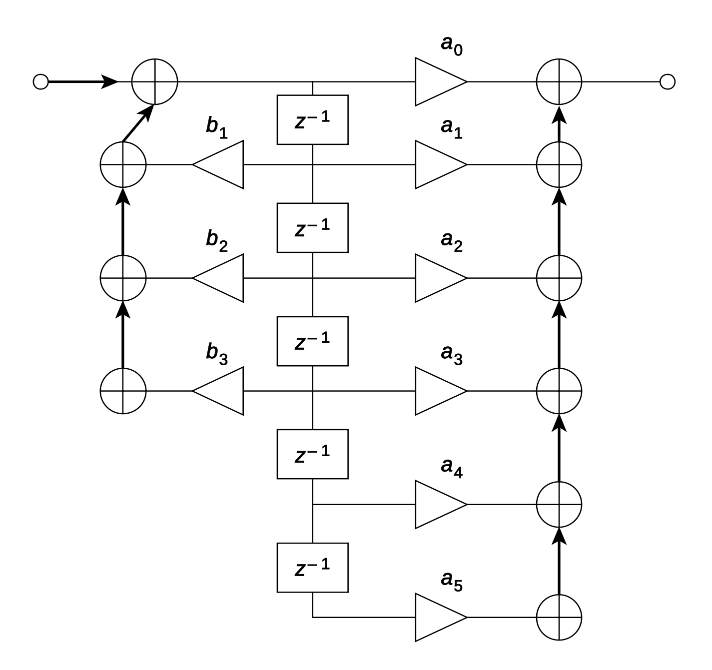


FIG. 12.5. *Structure d'un élément polyphase de synthèse*

La méthode de calcul des éléments polyphases de synthèse exposée ci-dessus est générale et s'applique à tout filtre d'analyse, à la seule condition qu'il ne possède pas de zéro sur le cercle unité. Elle nécessite d'effectuer un calcul par branche, avec des coefficients obtenus différents. Elle permet de spécifier le filtre d'analyse relativement indépendamment du filtre de synthèse. Cependant, il peut être intéressant de sacrifier un peu de flexibilité pour obtenir un calcul plus systématique et plus simple, comme au chapitre précédent.

12.4 BANCS DE FILTRES PSEUDO-QMF

Le principe repose sur l'hypothèse que, pour un filtre donné, l'affaiblissement est tel que les repliements ne proviennent que des bandes voisines. [3]

Soit H(Z) la fonction de transfert d'un filtre prototype passe-bas à phase linéaire ayant la réponse en fréquence représentée à la figure 12.6.

En considérant un nombre de coefficients égal à LN, il vient :

$$H(Z) = \sum_{k=0}^{LN-1} h_k Z^{-k} \tag{12.18}$$

Dans le banc, le filtre d'indice i, centré sur la fréquence $\frac{2i+1}{4N}$, a pour fonction de transfert $H\left(Ze^{-j2\pi\frac{2i+1}{4N}}\right)$.

Lors de l'analyse, une composante du signal située à la fréquence $\frac{2i+1}{4N} + \Delta f$, avec, par exemple, $\frac{1}{4N} < \Delta f < \frac{3}{4N}$, va être affaiblie suivant le facteur $H(\Delta f)$.

Le sous-échantillonnage à la cadence $\frac{1}{N}$ va produire un repliement de cette composante à la fréquence :

$$\frac{2i+1}{4N} + \frac{3}{4N} - \left(\frac{2i+1}{4N} + \Delta f\right) = \frac{3}{4N} - \Delta f.$$

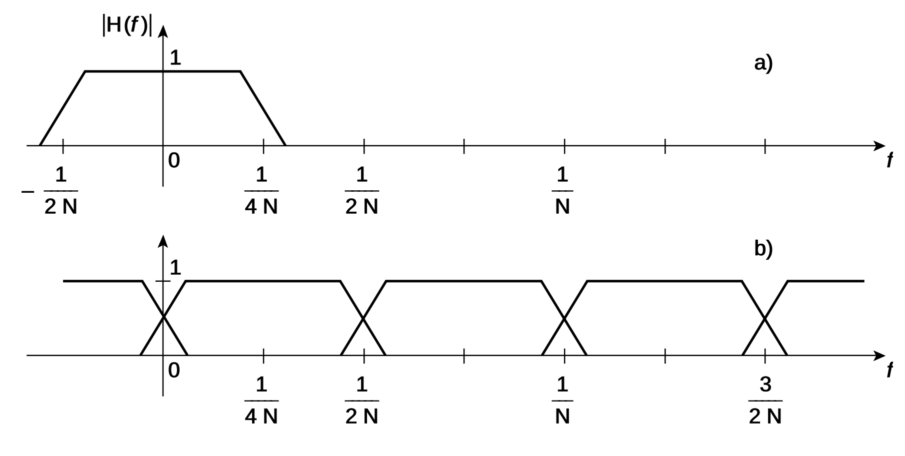


FIG. 12.6. *Banc de filtres réels*

a) Filtre prototype
b) Banc uniforme de N filtres réels

Lors de la synthèse, cette composante, repliée dans la bande du filtre d'indice i, va se trouver à la fréquence $\frac{2i+1}{4N} + \frac{1}{2N} - \Delta f$ et elle va subir l'affaiblissement du filtre de synthèse d'indice i, soit $G\left(\frac{1}{2N} - \Delta f\right)$, si G(Z) désigne le filtre prototype de synthèse. Finalement, la composante repliée aura subi l'affaiblissement :

$$H(\Delta f)\, G\left(\frac{1}{2N} - \Delta f\right).$$

Or, la même composante de signal va être traitée par le filtre d'indice $i + 1$, puisqu'elle tombe dans sa bande passante. L'échantillonnage produit ensuite une composante image, qui, lors de la synthèse, vient s'ajouter à la composante repliée précédente avec l'affaiblissement $H\left(\frac{1}{2N} - \Delta f\right) G(\Delta f)$. Le processus est illustré par la figure 12.7.

D'où la condition pour que ces composantes se compensent :

$$\left[H(\Delta f)\, G\left(\frac{1}{2N} - \Delta f\right)\right]_i + \left[H\left(\frac{1}{2N} - \Delta f\right) G(\Delta f)\right]_{i+1} = 0 \qquad (12.19)$$

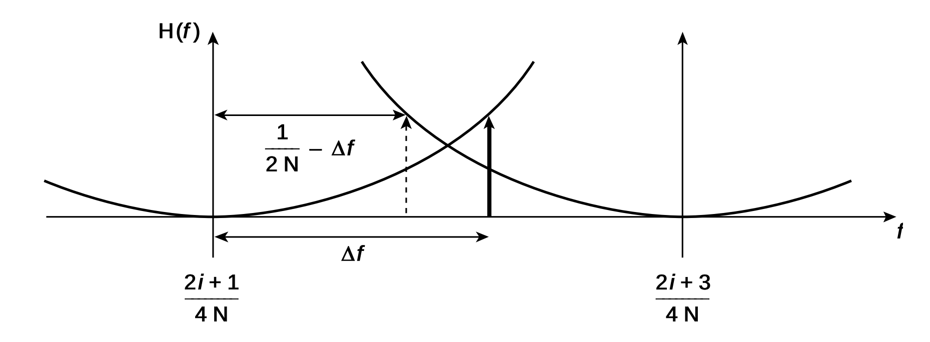


Fig. 12.7. *Repliement d'une composante dans le banc de filtres*

Cette condition d'absence de repliement peut être obtenue en prenant $G(f) = H(f)$ et en appliquant une différence de phase de $\frac{\pi}{2}$ entre les filtres d'indice i et $i + 1$, à l'analyse et à la synthèse.

La différence de phase nécessaire peut être obtenue en introduisant des déphasages dans les fonctions de modulation, par exemple en prenant pour les coefficients h_{ik} du filtre d'analyse d'indice i :

$$h_{ik} = 2h_k \cos\left[\left((2i+1)\frac{\pi}{2\mathrm{N}}\right)\left(k - \frac{\mathrm{LN}-1}{2}\right) + \theta_i\right] \qquad (12.20)$$

avec $0 \leqslant i \leqslant \mathrm{N}-1$ et $0 \leqslant k \leqslant \mathrm{NL}-1$.

Et pour les coefficients du filtre de synthèse :

$$g_{ik} = 2h_k \cos\left[\left((2i+1)\frac{\pi}{2\mathrm{N}}\right)\left(k - \frac{\mathrm{LN}-1}{2}\right) - \theta_i\right] \qquad (12.21)$$

En posant :

$$a_i = e^{j\theta_i}; \quad \mathrm{C}_i = e^{-j(2i+1)\frac{\pi}{2\mathrm{N}}\frac{\mathrm{LN}-1}{2}} \qquad (12.22)$$

il vient, pour les fonctions de transfert correspondantes :

$$\mathrm{H}_i(\mathrm{Z}) = a_i c_i \mathrm{H}\left(\mathrm{Z}e^{-j(2i+1)\frac{\pi}{2\mathrm{N}}}\right) + \bar{a}_i \bar{c}_i \mathrm{H}\left(\mathrm{Z}e^{j(2i+1)\frac{\pi}{2\mathrm{N}}}\right) \qquad (12.23)$$

$$\mathrm{G}_i(\mathrm{Z}) = \bar{a}_i c_i \mathrm{H}\left(\mathrm{Z}e^{-j(2i+1)\frac{\pi}{2\mathrm{N}}}\right) + a_i \bar{c}_i \mathrm{H}\left(\mathrm{Z}e^{j(2i+1)\frac{\pi}{2\mathrm{N}}}\right) \qquad (12.24)$$

Avec la symétrie des coefficients h_k et le centrage des fonctions de modulation, les relations suivantes sont vérifiées :

$$g_{ik} = h_{i(\mathrm{LN}-1-k)}; \quad \mathrm{G}_i(\mathrm{Z}) = \mathrm{Z}^{-(\mathrm{LN}-1)}\,\mathrm{H}_i(\mathrm{Z}^{-1}) \qquad (12.25)$$

Dans ces conditions, la réponse totale du système d'analyse et synthèse s'écrit :

$$\mathrm{T}(\mathrm{Z}) = \frac{\hat{\mathrm{X}}(\mathrm{Z})}{\mathrm{X}(\mathrm{Z})} = \frac{1}{\mathrm{N}} \sum_{i=0}^{\mathrm{N}-1} \mathrm{H}_i(\mathrm{Z})\,\mathrm{G}_i(\mathrm{Z})$$

$$= \frac{\mathrm{Z}^{-(\mathrm{LN}-1)}}{\mathrm{N}} \sum_{i=0}^{\mathrm{N}-1} \mathrm{H}_i(\mathrm{Z})\,\mathrm{H}_i(\mathrm{Z}^{-1}) \qquad (12.26)$$

C'est-à-dire que le système global est à phase linéaire.

Avec l'hypothèse que les filtres ont un affaiblissement suffisant pour que seules les bandes voisines puissent apporter un repliement significatif, il faut maintenant déterminer les valeurs à donner aux angles θ_i pour obtenir l'annulation désirée.

En sortie du filtre $\mathrm{H}_i(\mathrm{Z})$, après sous-échantillonnage, le signal s'écrit, conformément à la relation (X.8) :

$$\mathrm{X}_i(\mathrm{Z}^{\mathrm{N}}) = \frac{1}{\mathrm{N}} \sum_{m=0}^{\mathrm{N}-1} \mathrm{H}_i(\mathrm{ZW}^m)\,\mathrm{X}(\mathrm{ZW}^m) \qquad (12.27)$$

Et pour le signal en sortie du système, il vient :

$$\hat{\mathrm{X}}(\mathrm{Z}) = \sum_{i=0}^{\mathrm{N}-1} \mathrm{G}_i(\mathrm{Z})\,\mathrm{X}_i(\mathrm{Z}_{\mathrm{N}}) = \frac{1}{\mathrm{N}} \sum_{m=0}^{\mathrm{N}-1} \mathrm{X}(\mathrm{ZW}^m) \sum_{i=0}^{\mathrm{N}-1} \mathrm{G}_i(\mathrm{Z})\,\mathrm{H}_i(\mathrm{ZW}^m) \qquad (12.28)$$

La condition de parfaite reconstitution s'écrit alors :

$$\sum_{i=0}^{N-1} G_i(Z)\, H_i(Z) = Z^{-k} \tag{12.29}$$

$$\sum_{i=0}^{N-1} G_i(Z)\, H_i(ZW^m) = 0; \quad 1 \leqslant m \leqslant N-1 \tag{12.30}$$

Pour faire apparaître l'annulation des composantes repliées, il faut examiner le signal en sortie de chacun des filtres de synthèse. En sortie du filtre $G_i(Z)$, le signal $\hat{X}_i(Z)$ peut s'écrire, en reprenant la relation (12.27) :

$$\hat{X}_i(Z) = G_i(Z)\, \frac{1}{N} \sum_{m=0}^{N-1} H_i(ZW^m)\, X(ZW^m) \tag{12.31}$$

Mais, d'après la définition (12.24) et les hypothèses faites sur l'affaiblissement, le filtre $G_i(Z)$ ne laisse passer que la bande centrée sur la fréquence $\frac{2i+1}{4N}$ et les deux bandes voisines. Compte tenu de la répartition des bandes sur l'axe des fréquences, les indices m associés à ces bandes voisines correspondent à des translations de fréquence telles que :

$$\frac{2i+1}{4M} \pm \frac{m}{N} = -\frac{2i+1}{4M} \pm \frac{1}{2N} \tag{12.32}$$

En effet, le sous-échantillonnage amène des translations de fréquence qui sont des multiples entiers de la fréquence $\frac{1}{N}$.

Dans ces conditions, les valeurs de m s'écrivent :

$$m = \pm\, i \quad \text{et} \quad m = \pm\,(i+1)$$

Par exemple, en reprenant le cas de la figure 12.7, la composante repliée provient d'une composante à la fréquence $-\left(\frac{2i+1}{4N} + \Delta f\right)$ décalée de $\frac{i+1}{N}$, soit :

$$-\left(\frac{2i+1}{4N} + \Delta f\right) + \frac{i+1}{N} = \frac{2i+3}{4N} - \Delta f$$

Par suite $\hat{X}_i(Z)$ se limite au développement suivant, en reprenant la définition (12.23) de $H_i(Z)$:

$$\hat{X}_i(Z) = G_i(Z)\, \frac{1}{N} \left[a_i c_i H\left(ZW^{-\frac{2i+1}{4}}\right)\right] X(Z) + \bar{a}_i \bar{c}_i H\left(ZW^{\frac{2i+1}{4}}\right) X(Z)$$

$$+ a_i c_i \mathrm{H}\left(\mathrm{ZW}^{\frac{2i-1}{4}}\right) \mathrm{X}(\mathrm{ZW}^{i}) + \bar{a}_i \bar{c}_i \mathrm{H}\left(\mathrm{ZW}^{\frac{1-2i}{4}}\right) \mathrm{X}(\mathrm{ZW}^{-i})$$

$$+ a_i c_i \mathrm{H}\left(\mathrm{ZW}^{\frac{2i+3}{4}}\right) \mathrm{X}(\mathrm{ZW}^{i+1}) + \bar{a}_i \bar{c}_i \mathrm{H}\left(\mathrm{ZW}^{-\frac{2i+3}{4}}\right) \mathrm{X}[\mathrm{ZW}^{-(i+1)}] \quad (12.33)$$

Comme :

$$\hat{\mathrm{X}}(\mathrm{Z}) = \sum_{i=0}^{\mathrm{N}-1} \hat{\mathrm{X}}_i(\mathrm{Z}) \quad (12.34)$$

les repliements s'annulent dans $\hat{\mathrm{X}}(\mathrm{Z})$ si la bande haute de $\hat{\mathrm{X}}_i(\mathrm{Z})$ compense la bande basse de $\hat{\mathrm{X}}_{i+1}(\mathrm{Z})$. La condition correspondante s'obtient en reportant dans l'expression de $\hat{\mathrm{X}}_i(\mathrm{Z})$ la définition (12.24) de $\mathrm{G}_i(\mathrm{Z})$ et en écrivant la même expression pour $\hat{\mathrm{X}}_{i+1}(\mathrm{Z})$. Les facteurs de $\mathrm{X}[\mathrm{ZW}^{i+1}]$ et $\mathrm{X}[\mathrm{ZW}^{-(i+1)}]$ s'annulent si la condition suivante est vérifiée :

$$a_i^2 c_i \bar{c}_i \mathrm{H}\left(\mathrm{ZW}^{\frac{2i+3}{4}}\right) \mathrm{H}\left(\mathrm{ZW}^{\frac{2i+1}{4}}\right) + a_{i+1}^2 c_{i+1} \bar{c}_{i+1} \mathrm{H}\left(\mathrm{ZW}^{\frac{2i+1}{4}}\right) \mathrm{H}\left(\mathrm{ZW}^{\frac{2i+3}{4}}\right) = 0$$

c'est-à-dire si :

$$a_{i+1}^2 = -a_i^2 \quad (12.35)$$

Il faut donc que les déphasages vérifient :

$$\theta_{i+1} = \theta_i + \frac{\pi}{2}\,; \quad 0 \leqslant i \leqslant \mathrm{N}-1 \quad (12.36)$$

La première condition de parfaite reconstitution (12.29) s'écrit alors :

$$\frac{1}{\mathrm{N}} \sum_{i=0}^{\mathrm{N}-1} \left[c_i \mathrm{H}\left(\mathrm{ZW}^{\frac{2i+1}{4}}\right)\right]^2 + \left[\bar{c}_i \mathrm{H}\left(\mathrm{ZW}^{-\frac{2i+1}{4}}\right)\right] + (a_0^2 + \bar{a}_0^2)\, \mathrm{H}\left(\mathrm{ZW}^{\frac{1}{4}}\right) \mathrm{H}\left(\mathrm{ZW}^{-\frac{1}{4}}\right)$$

$$+ (a_{\mathrm{N}-1}^2 + \bar{a}_{\mathrm{N}-1}^2)\, \mathrm{H}\left(\mathrm{ZW}^{\frac{\mathrm{N}-1}{4}}\right) \mathrm{H}\left(\mathrm{ZW}^{-\frac{\mathrm{N}-1}{4}}\right) = 1 \quad (12.37)$$

car les produits croisés ne s'annulent pas à l'origine et à la demi-fréquence d'échantillonnage puisque les filtres sont voisins.

Pour faire disparaître ces produits croisés, il faut prendre :

$$\theta_i = (-1)^i \frac{\pi}{4} \quad (12.38)$$

et la condition imposée pour le calcul de la réponse en fréquence du filtre prototype s'écrit finalement :

$$|\mathrm{H}(f)|^2 + \left|\mathrm{H}\left(\frac{1}{2\mathrm{N}} - f\right)\right|^2 = 1\,; \quad 0 \leqslant f < \frac{1}{2\mathrm{N}} \quad (12.39)$$

$$H(f) = 0\,; \qquad \frac{1}{2N} \leqslant f \leqslant \frac{1}{2}$$

A noter que les choix suivants sont également possibles pour les déphasages dans les bancs d'analyse et de synthèse :

- $\theta_i = (i+1)\,\dfrac{\pi}{2}$; la réponse totale du système s'annule aux fréquences 0 et $\dfrac{1}{2}$.
- $\theta_i = i\,\dfrac{\pi}{2}$; la réponse totale double aux fréquences 0 et $\dfrac{1}{2}$.

En résumé, la procédure de calcul d'un banc de N filtres réels pseudo-QMF comprend les deux opérations suivantes :

1. Calcul du filtre prototype à phase linéaire approchant les spécifications (12.39), avec LN coefficients.

2. Détermination des fonctions de transfert des filtres d'analyse et synthèse par les expressions :

$$H_i(Z) = 2 \sum_{k=0}^{LN-1} \cos\left[2\pi\, \frac{2i+1}{4N}\left(k - \frac{LN-1}{2}\right) + (2i+1)\,\frac{\pi}{4}\right] h_k Z^{-k} \qquad (12.40)$$

$$G_i(Z) = 2 \sum_{k=0}^{LN-1} \cos\left[2\pi\, \frac{2i+1}{4N}\left(k - \frac{LN-1}{2}\right) - (2i+1)\,\frac{\pi}{4}\right] h_k Z^{-k} \qquad (12.41)$$

Le gabarit du filtre prototype reflète la séparation désirée entre les sous-bandes. Plusieurs approches peuvent être envisagées pour le calcul.

12.5 CALCUL DES COEFFICIENTS DU FILTRE PROTOTYPE

C'est le calcul des coefficients d'un filtre demi-Nyquist et la première approche consiste à prendre une bande de transition en cosinus et à utiliser la formule (5.37). Une autre approche, plus efficace, consiste à utiliser la méthode de l'échantillonnage en fréquence évoquée au paragraphe 5.4. En particulier, quand la bande de transition est égale à l'espacement entre les sous-bandes, les coefficients sont obtenus par une formule simple. [4].

Soit K un entier et un ensemble de KN échantillons H_k $(0 \leqslant k \leqslant KN-1)$ dans le domaine des fréquences, tels que :

$$H_0 = 1 \qquad (12.42)$$

$$H_k^2 + H_{K-k}^2 = 1\,; \quad H_{KN-k} = H_k\,; \quad 1 \leqslant k \leqslant K-1$$

$$H_k = 0\,; \quad K \leqslant k \leqslant KN-K$$

Généralement le nombre N de sous-bandes est pair. Les coefficients du filtre correspondant h_i $(0 \leqslant i \leqslant KN - 1)$ sont obtenus par TFD inverse.

La relation :

$$H_0 + 2 \sum_{k=1}^{K-1} (-1)^k H_k = 0 \tag{12.43}$$

permet d'annuler le coefficient milieu $h_{KN/2}$, ce qui amène un nombre impair de coefficients pour le filtre. Il en résulte un zéro double à la demi-fréquence d'échantillonnage, comme indiqué au paragraphe V.12 et donc un affaiblissement important aux fréquences élevées.

Pour K = 3 et K = 4, les équations (12.42-43) définissent un système déterminé et les échantillons en fréquence prennent les valeurs :

$$H_1 = 0{,}911438\,; \quad H_2 = 0{,}411438$$

et

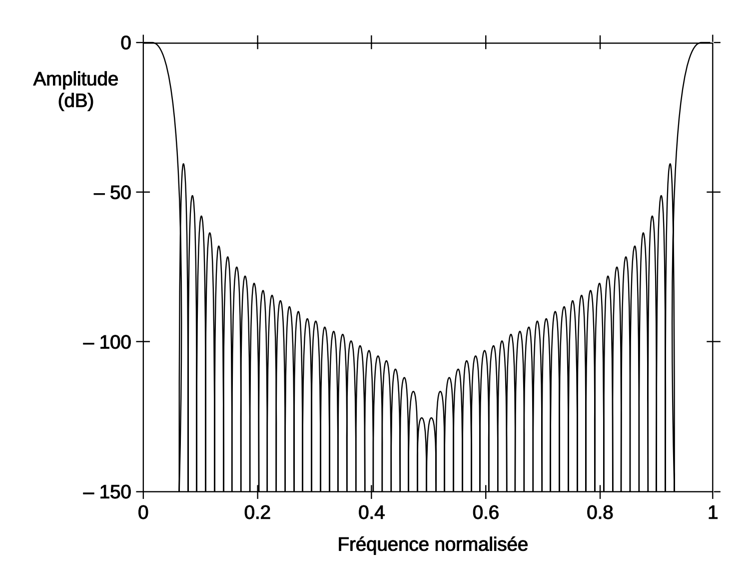
$$H_1 = 0{,}971960\,; \quad H_2 = \frac{1}{\sqrt{2}}\,; \quad H_3 = 0{,}235147 \tag{12.44}$$

Par exemple, pour un banc de N = 16 filtres avec K = 4, les coefficients sont donnés par la relation :

$$h_{i+1} = 1 + 2 \sum_{k=1}^{K-1} (-1)^k H_k \cos(2\pi ki/KN)\,; \quad 1 \leqslant i \leqslant 63 \tag{12.45}$$

$$h_1 = 0.$$

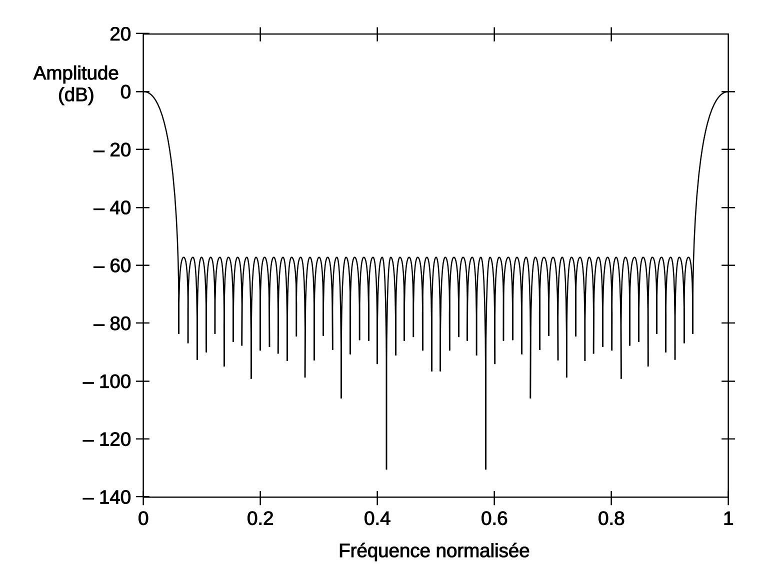
Le filtre obtenu est à nombre impair de coefficients. La réponse en fréquence, dont la bande de transition est égale à 1/16, est donnée à la figure (12.8), avec les valeurs des 63 coefficients (h_1 est nul).



h_i	Coefficients	h_i	Coefficients
h_1	1,9999999e-09	$h_{17} = h_{49}$	– 4,1421356e-01
$h_2 = h_{64}$	2,4374869e-03	$h_{18} = h_{48}$	– 3,3302151e-01
$h_3 = h_{63}$	8,9600890e-03	$h_{19} = h_{47}$	– 1,8860434e-01
$h_4 = h_{62}$	1,7309178e-02	$h_{20} = h_{46}$	2,4872388e-02
$h_5 = h_{61}$	2,4078729e-02	$h_{21} = h_{45}$	3,0941094e-01
$h_6 = h_{60}$	2,5214011e-02	$h_{22} = h_{44}$	6,6263312e-01
$h_7 = h_{59}$	1,6635762e-02	$h_{23} = h_{43}$	1,0775305e+00
$h_8 = h_{58}$	– 5,0758287e-03	$h_{24} = h_{42}$	1,5425480e+00
$h_9 = h_{57}$	– 4,2010768e-02	$h_{25} = h_{41}$	2,0420108e+00
$h_{10} = h_{56}$	– 9,4346766e-02	$h_{26} = h_{40}$	2,5568746e+00
$h_{11} = h_{55}$	– 1,5992266e-01	$h_{27} = h_{39}$	3,0657564e+00
$h_{12} = h_{54}$	– 2,3402304e-01	$h_{28} = h_{38}$	3,5461759e+00
$h_{13} = h_{53}$	– 3,0941094e-01	$h_{29} = h_{37}$	3,9759213e+00
$h_{14} = h_{52}$	– 3,7662359e-01	$h_{30} = h_{36}$	4,3344420e+00
$h_{15} = h_{51}$	– 4,2452159e-01	$h_{31} = h_{35}$	4,6041658e+00
$h_{16} = h_{50}$	– 4,4105818e-01	$h_{32} = h_{34}$	4,7716422e+00
		h_{33}	4,8284271e+00

FIG. 12.8. *Filtre prototype obtenu par échantillonnage en fréquence, pour banc de 16 filtres et 64 coefficients*

La figure montre que l'affaiblissement croît avec la fréquence. Il est aussi possible d'obtenir un affaiblissement constant en reprenant la méthode du paragraphe 11.2 avec des spécifications adaptées. Des techniques itératives peuvent compléter la procédure pour mieux approcher la condition de symétrie (12.39). Un exemple de filtre est donné à la figure 12.9. Il possède 64 coefficients, fournit un affaiblissement de 58 dB et sa bande de transition est de 1/16, comme à la figure 12.8.



h_i	Coefficients	h_i	Coefficients
$h_1 = h_{64}$	9,2839870e-004	$h_{17} = h_{48}$	– 6,4021587e-003
$h_2 = h_{63}$	5,9685008e-004	$h_{18} = h_{47}$	– 4,7447370e-003
$h_3 = h_{62}$	6,8568814e-004	$h_{19} = h_{46}$	– 2,1424791e-003
$h_4 = h_{61}$	6,8789993e-004	$h_{20} = h_{45}$	1,4680266e-003
$h_5 = h_{60}$	5,6482087e-004	$h_{21} = h_{44}$	6,0982460e-003
$h_6 = h_{59}$	2,8097967e-004	$h_{22} = h_{43}$	1,1699623e-002
$h_7 = h_{58}$	– 1,9044096e-004	$h_{23} = h_{42}$	1,8159361e-002
$h_8 = h_{57}$	– 8,6214757e-004	$h_{24} = h_{41}$	2,5302338e-002
$h_9 = h_{56}$	– 1,7275867e-003	$h_{25} = h_{40}$	3,2898876e-002
$h_{10} = h_{55}$	– 2,7557760e-003	$h_{26} = h_{39}$	4,0672058e-002
$h_{11} = h_{54}$	– 3,8879464e-003	$h_{27} = h_{38}$	4,8309834e-002
$h_{12} = h_{53}$	– 5,0372397e-003	$h_{28} = h_{37}$	5,5491156e-002
$h_{13} = h_{52}$	– 6,0890424e-003	$h_{29} = h_{36}$	6,1893832e-002
$h_{14} = h_{51}$	– 6,9067117e-003	$h_{30} = h_{35}$	6,7223291e-002
$h_{15} = h_{50}$	– 7,3368967e-003	$h_{31} = h_{34}$	7,1226098e-002
$h_{16} = h_{49}$	– 7,2203731e-003	$h_{32} = h_{33}$	7,3708998e-002

FIG. 12.9. *Filtre prototype pour banc de 16 filtres*

En normalisant par le facteur 64 les valeurs des coefficients du tableau de la figure 12.8, on vérifie que les deux techniques de calcul conduisent à des valeurs proches.

Une fois les coefficients calculés, il faut procéder à la mise en œuvre et agencer les calculs pour que la quantité d'opérations arithmétiques soit minimale.

12.6 RÉALISATION D'UN BANC DE FILTRES RÉELS

Soit à réaliser un banc de N filtres réels ayant les réponses en fréquence de la figure 12.6 et un nombre pair de coefficients 2LN, dans lequel le filtre d'indice i a pour coefficients :

$$h_{ik} = h_k \cos\left[\frac{2\pi}{2N}\left(i+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right] \qquad (12.46)$$

avec $0 \leqslant i \leqslant N$; $0 \leqslant k \leqslant 2LN - 1$.

Une décomposition en réseau polyphasé et transformée de Fourier Discrète peut être obtenue en posant $k = 2Nl + m$, avec $0 \leqslant l \leqslant L - 1$ et $0 \leqslant m \leqslant 2N - 1$.

En effet, la sortie $x_i(n)$ du filtre d'indice i s'écrit :

$$x_i(n) = \sum_{k=0}^{2LN-1} x(n-k)\, h_k \cos\left[\frac{2\pi}{2N}\left(i+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right] \qquad (12.47)$$

ou encore, en remplaçant k par $2Nl + m$ et en simplifiant :

$$x_i(n) = \sum_{m=0}^{2N-1} \cos\frac{2\pi}{2N}\left(i+\frac{1}{2}\right)\left(m+\frac{1}{2}\right) \sum_{l=0}^{L-1} (-1)^l\, h_{2Nl+m}\, x(n-2Nl-m) \qquad (12.48)$$

En appliquant au filtre prototype la décomposition générale (X.36) des fonctions de transfert, il vient :

$$H(Z) = \sum_{m=0}^{2N-1} Z^{-m}\, H_m(Z^{2N}) \qquad (12.49)$$

et les filtres $H_m(Z^{2N})$ sont ceux qui interviennent dans la seconde sommation de l'expression (12.48). Pour tenir compte du facteur $(-1)^l$, il suffit d'introduire les fonctions $H_m(-Z^{2N})$ et le schéma correspondant aux filtres d'analyse est donné à la figure 12.10.

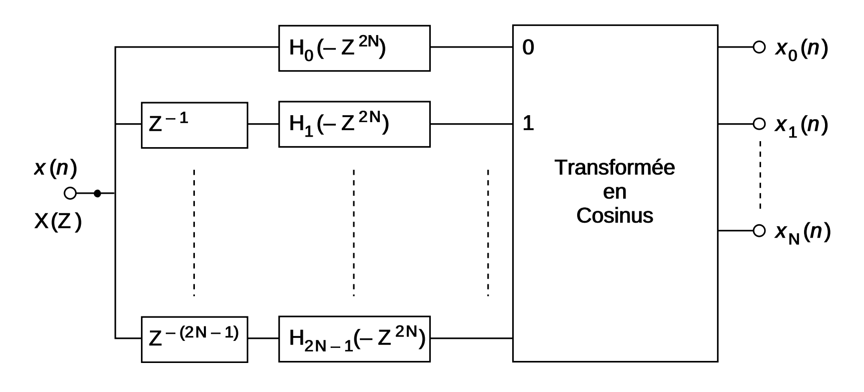


FIG. 12.10. *Schéma d'un banc de N filtres réels*

La décomposition fait apparaître un réseau polyphasé à 2N branches et une transformée en cosinus.

Ce schéma peut encore se simplifier, car, dans la transformation en cosinus considérée, deux entrées symétriques subissent les mêmes opérations, au signe près. En factorisant, on aboutit à la transformée doublement impaire en cosinus qui est un cas particulier mentionné au paragraphe III.5.2. De plus, avec le sous-échantillonnage, le système fonctionne à la cadence $\frac{1}{N}$. Comme l'ensemble comporte 2N branches, une donnée $x(n)$ se trouve traitée par les filtres H_i et H_{i+N} à deux instants successifs. Dans ces conditions, on peut regrouper les 2N branches du réseau polyphasé en $\frac{N}{2}$ sous-ensembles ayant la structure en treillis de la figure 12.11. Le schéma global est alors celui de la figure 12.12. Dans le cas des filtres pseudo QMF étudiés au paragraphe précédent, ce schéma s'applique avec l'introduction des déphasages. En effet, en reprenant la relation (12.47) avec les coefficients de filtre de l'expression (12.40), la sortie $x_i(n)$ du filtre d'indice i à 2LN coefficients s'écrit :

$$x_i(n) = 2 \sum_{k=0}^{2LN-1} x(n-k)\, h_k \cos\left[2\pi \frac{2i+1}{4N}\left(k - \frac{2LN-1}{2}\right) + (2i+1)\frac{\pi}{4}\right] \quad (12.50)$$

En posant cette fois $k = lN + m$, on obtient la double sommation suivante :

$$x_i(n) = 2 \sum_{m=0}^{N-1} \sum_{l=0}^{2L-1} \cos\left[(2i+1)(2m+1+N)\frac{\pi}{4N} + (2i+1)(l-L)\frac{\pi}{2}\right] h_{lN+m}\, x(n-lN-m) \quad (12.51)$$

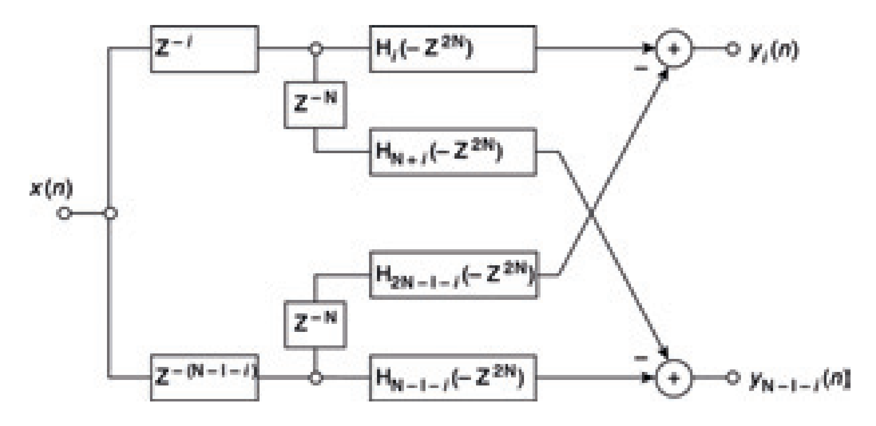


FIG. 12.11. *Structure en treillis pour le réseau polyphasé*

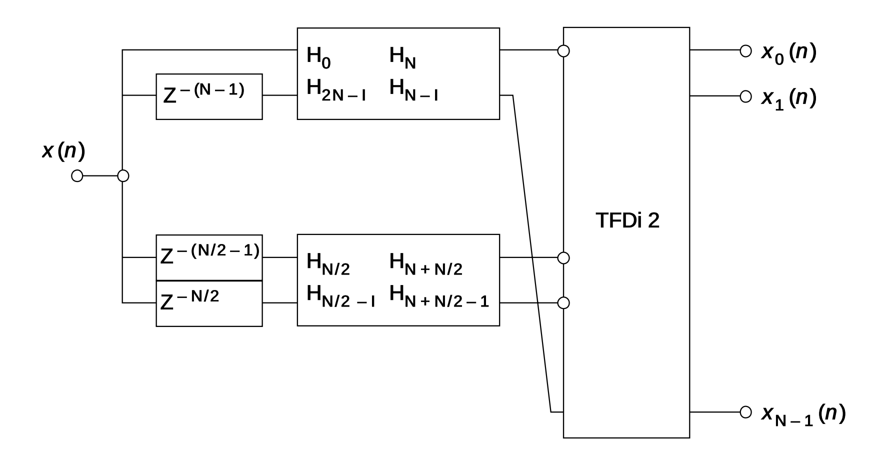


FIG. 12.12. *Structure optimisée du banc de N filtres réels*

Le développement des cosinus fait apparaître les termes suivants :

$$\cos\,(2i+1)\,(l-\mathrm{L})\,\frac{\pi}{2} = \cos\,(l-\mathrm{L})\,\frac{\pi}{2}$$

et :

$$\sin\,(2i+1)\,(l-\mathrm{L})\,\frac{\pi}{2} = (-1)^i \sin\,(l-\mathrm{L})\,\frac{\pi}{2}$$

De plus, la relation suivante est vérifiée :

$$\cos\left[(2i+1)\,[2\,(\mathrm{N}-1-m)+1]\,\frac{\pi}{4\mathrm{N}}\right] = (-1)^i \sin\left[(2i+1)\,(2m+1)\,\frac{\pi}{4\mathrm{N}}\right]$$

Enfin, pour tenir compte du terme N dans l'argument du cosinus de (12.51), il suffit de décaler les données en indice de $\frac{\mathrm{N}}{2}$. En combinant tous ces résultats, la sortie $x_i(n)$ se calcule comme suit :

$$x_i(n) = 2 \sum_{m=0}^{\mathrm{N}-1} \cos\left[(2i+1)\,(2m+1)\,\frac{\pi}{4\mathrm{N}}\right] y_m(n) \qquad (12.52)$$

avec :

$$y_m(n) = \left[-y_{2,\,\frac{\mathrm{N}}{2}-1-m}(n) + y_{2,\,\frac{\mathrm{N}}{2}+m}(n)\right]; \quad 0 \leqslant m \leqslant \frac{\mathrm{N}}{2} - 1$$

$$y_m(n) = \left[y_{1,\,m-\frac{\mathrm{N}}{2}}(n) - y_{1,\,3\frac{\mathrm{N}}{2}-m-1}(n)\right]; \quad \frac{\mathrm{N}}{2} \leqslant m \leqslant \mathrm{N}-1$$

et :

$$y_{1,m}(n) = \sum_{m=0}^{2\mathrm{L}-1} \cos\,(l-\mathrm{L})\,\frac{\pi}{2}\,h_{l\mathrm{N}+m}\,x(n-l\mathrm{N}-m)$$

$$y_{2,m}(n) = \sum_{m=0}^{2\mathrm{L}-1} \sin\,(l-\mathrm{L})\,\frac{\pi}{2}\,h_{l\mathrm{N}+m}\,x(n-l\mathrm{N}-m)$$

Les suites $y_{1,m}(n)$ et $y_{2,m}(n)$ sont entrelacées, avec la fréquence d'échantillonnage $\frac{1}{2\mathrm{N}}$ et le banc de filtres d'analyse se réalise à l'aide d'une transformée en cosinus doublement impaire. Finalement, les déphasages introduits par la technique pseudo-QMF ont été pris en compte simplement par des réarrangements de données avant la transformée.

BIBLIOGRAPHIE

[1] M. Bellanger and J. Daguet – «TDM-FDM Transmultiplexer : Digital Polyphase and FFT», *IEEE Trans. on Communications,* Vol. COM-22, n° 9, Sept. 1974, pp. 1199-1205.

[2] R. E. Crochière and L. R. Rabiner – *Multirate Digital Signal Processing*, Prentice-Hall Inc., Englewood Cliffs, N. J., 1983.

[3] N. J. Fliege – *Multirate Digital Signal Processing*, John Wiley, Chichester, 1994.

[4] K. W. Martin – «Small Side-lobe Filter Design for Multitone Data Communications», IEEE Trans – CAS II, Vol. 45, N° 8, August 1998, pp. 1155-1161.

Chapitre 13

Analyse et modélisation

La modélisation des systèmes est l'un des grands domaines du traitement du signal. Par ailleurs, la modélisation des signaux constitue une autre approche pour leur analyse, avec des propriétés qui diffèrent de celles de la transformée de Fourier et des filtres définis dans le domaine des fréquences. La prédiction linéaire, en particulier, est un outil simple et efficace pour caractériser certains type de signaux et procéder à leur compression. Les traitements sont spécifiés dans le domaine temporel, en utilisant les paramètres statistiques et principalement la corrélation.

13.1 Autocorrélation et intercorrélation

Pour mesurer le degré de similitude de deux signaux on définit un coefficient de corrélation. Il est naturel de faire correspondre la valeur 1 de ce coefficient à deux signaux identiques, la valeur zéro à deux signaux sans aucune relation et la valeur – 1 à des signaux opposés. Quand on compare des signaux décalés dans le temps, le coefficient de corrélation devient une fonction de temps, et l'on obtient la fonction d'intercorrélation si les signaux sont différents et la fonction d'autocorrélation s'ils sont identiques.

Des définitions et des propriétés des signaux aléatoires vont maintenant être rappelées, pour reprendre et compléter les paragraphes 1.8 et 4.4.

Comme indiqué au paragraphe 1.8, la fonction d'autocorrélation du signal aléatoire discret $x(n)$ est la suite $r_{xx}(p)$ telle que :

$$r_{xx}(p) = \mathrm{E}[x(i).x(i-p)] \tag{13.1}$$

où $\mathrm{E}[x]$ désigne l'espérance de x.

Avec l'hypothèse d'ergodicité, il vient :

$$r_{xx}(p) = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{i=-N}^{N} x(i).x(i-p) \tag{13.2}$$

La fonction $r_{xx}(p)$ est paire ; elle a comme valeur à l'origine la puissance du signal et quel que soit n :

$$|r_{xx}(p)| \leqslant r_{xx}(0) \tag{13.3}$$

Soit un ensemble de N coefficients $h_i (1 \leqslant i \leqslant \mathrm{N})$. Le calcul de la variance de la variable $y(n)$ telle que :

$$y(n) = \sum_{i=1}^{\mathrm{N}} h_i x(n-i)$$

conduit à :

$$\mathrm{E}[y^2(n)] = \sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{N}} h_i h_j \mathrm{E}[x(n-i).x(n-j)]$$

soit :

$$\mathrm{E}[y^2(n)] = \sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{N}} h_i h_j r_{xx}(i-j) \tag{13.4}$$

Comme cette variance est positive ou nulle, il en résulte l'inégalité suivante :

$$\sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{N}} h_i h_j r_{xx}(i-j) \geqslant 0 \tag{13.5}$$

Cette propriété caractérise les fonctions de type positif.

Si dans la définition (13.1) on remplace $x(i-n)$ par un autre signal, on obtient une fonction qui permet de comparer deux signaux différents décalés. La fonction d'intercorrélation entre deux signaux discrets $x(n)$ et $y(n)$ est la suite $r_{xy}(p)$ telle que :

$$r_{xy}(p) = \mathrm{E}[x(i)y(i-p)] \tag{13.6}$$

Avec l'hypothèse d'ergodicité :

$$r_{xy}(p) = \lim_{\mathrm{N}\to\infty} \frac{1}{2\mathrm{N}+1} \sum_{i=-\mathrm{N}}^{\mathrm{N}} x(i)y(i-p) \tag{13.7}$$

On a également :

$$r_{xy}(-p) = \mathrm{E}[x(i)y(i+p)] = r_{yx}(p) \tag{13.8}$$

Par exemple dans le cas où ces signaux constituent l'entrée et la sortie d'un filtre de coefficients h_m :

$$y(n) = \sum_{m=0}^{\infty} h_m x(n-m)$$

il vient, comme indiqué au paragraphe 4.4 :

$$r_{yx}(p) = \mathrm{E}[y(i).x(i-p)] = \sum_{m=0}^{\infty} h_m r_{xx}(p-m)$$

soit :

$$r_{yx}(p) = r_{xx}(p) * h_m \qquad (13.9)$$

De même :

$$r_{xy}(p) = r_{xx}(p) * h_{-m} \qquad (13.10)$$

et également :

$$r_{yy}(p) = r_{xx}(p) * h_m * h_{-m} \qquad (13.11)$$

Si deux signaux aléatoires centrés sont indépendants, leurs fonctions d'intercorrélation sont nulles. De plus, on a toujours l'inégalité :

$$|r_{xy}(p)| \leqslant \frac{1}{2}[r_{xx}(0) + r_{yy}(0)] \qquad (13.12)$$

Il est intéressant de signaler que le calcul des fonctions d'autocorrélation et d'intercorrélation peut, dans certains cas, se faire sans multiplication, en remplaçant les signaux par leur signe. En particulier, si $x(n)$ est un signal gaussien, on montre que [2] :

$$r_{xx}(p) = \sqrt{\frac{\pi}{2}}.\sqrt{r_{xx}}(0).\mathrm{E}[x(i).\mathrm{signe}\,[x(i-p)]] \qquad (13.13)$$

$$r_{xx}(p) = r_{xx}(0).\sin\left(\frac{\pi}{2}.\mathrm{E}[\mathrm{signe}\,[x(i).x(i-p)]]\right) \qquad (13.14)$$

Ainsi toute l'information contenue dans un signal gaussien est fournie par ses passages par zéro. Ces relations, qui peuvent s'étendre à d'autres types de signaux, permettent de simplifier les calculs et, par suite, les matériels.
La transformée de Fourier $\Phi_{xy}(f)$ de la fonction d'intercorrélation $r_{xy}(p)$ prend le nom d'interspectre :

$$\Phi_{xy}(f) = \mathrm{X}(f).\overline{\mathrm{Y}(f)}$$

où $\mathrm{X}(f)$ désigne le spectre de la suite $x(n)$ et $\overline{\mathrm{Y}(f)}$ le spectre conjugué de la suite $y(n)$.

Si la suite $y(n)$ est la sortie d'un filtre de fonction de transfert $\mathrm{H}(f)$, il vient :

$$\mathrm{H}(f) = \frac{\mathrm{Y}(f)}{\mathrm{X}(f)} = \frac{\mathrm{Y}(f)\overline{\mathrm{X}(f)}}{\mathrm{X}(f)\overline{\mathrm{X}(f)}}$$

D'où la relation :

$$\Phi_{yx}(f) = \Phi_{xx}(f).\mathrm{H}(f) \qquad (13.15)$$

qui correspond à (13.9). De même à la relation (13.10) correspond :

$$\Phi_{xy}(f) = \Phi_{xx}(f).\overline{\mathrm{H}(f)}$$

et finalement :

$$\Phi_{yy}(f) = \Phi_{xx}(f).|\mathrm{H}(f)|^2 \qquad (13.16)$$

Ces résultats s'appliquent à l'analyse spectrale des signaux aléatoires en général et sont utiles pour l'étude des systèmes adaptatifs.

13.2 ANALYSE SPECTRALE PAR CORRÉLOGRAMME

La transformée de Fourier de la fonction d'autocorrélation est la densité spectrale de puissance du signal :

$$S(f) = \sum_{p=-\infty}^{\infty} r(p)e^{-j2\pi pf} \qquad (13.17)$$

en désignant par r(p) la fonction d'autocorrélation (AC) du signal. En pratique, l'analyse du signal se fait à partir d'un nombre limité, N_0, d'échantillons. Il faut donc commencer par estimer les valeurs de r(p).

Une première estimation de la fonction AC est fournie par l'expression :

$$r_1(p)= \frac{1}{N_0} \sum_{n=p+1}^{N_0} x(n)x(n-p) \qquad (13.18)$$

Elle est biaisée car :

$$E[r_1(p)] = \frac{N_0-p}{N_0} r(p)$$

En fait, pour avoir une estimation non biaisée il faut prendre :

$$r_2(p) = \frac{1}{N_0-p} \sum_{n=p+1}^{N_0} x(n)x(n-p) \qquad (13.19)$$

A partir de P valeurs de la fonction AC, l'estimation spectrale dite corrélogramme est donnée par :

$$S_{CR}(f) = \sum_{p=-(P-1)}^{P-1} r_2(p)e^{-j2\pi pf} \qquad (13.20)$$

ou encore, en fonction du spectre théorique :

$$S_{CR}(f) = S(f) * \frac{\sin \pi f(2P-1)}{\sin \pi f} \qquad (13.21)$$

La variance est approximativement donnée par :

$$\mathrm{var}\{S_{CR}(f)\} \approx \frac{(2P-1)}{N_0} S_2(f) \qquad (13.22)$$

Il apparaît alors qu'il faut prendre le minimum de valeurs de la fonction AC pour effectuer l'estimation, c'est à dire qu'il faut se limiter aux valeurs significatives de la fonction AC résultant de l'estimation [1].

L'approche la plus directe pour obtenir le spectre de puissance d'un signal à partir de N_0 échantillons consiste à calculer la transformée de Fourier discrète et à prendre les valeurs $\left|X(k)\right|^2$. Le développement de ce terme montre que c'est l'estimation biaisée (13.18) de la fonction AC qui intervient dans cette procédure.

13.3 MATRICE D'AUTOCORRÉLATION

La matrice d'autocorrélation (AC) de dimension N d'un signal est la matrice carrée suivante :

$$R_N = \begin{bmatrix} r(0) & r(1) & r(2) & \dots & r(N-1) \\ r(1) & r(0) & r(1) & \dots & r(N-2) \\ r(2) & r(1) & r(0) & \dots & r(N-3) \\ \vdots & \vdots & \vdots & & \vdots \\ r(N-1) & r(N-2) & r(N-3) & \dots & r(0) \end{bmatrix}$$

La fonction d'autocorrélation étant de type positif (13.5), la matrice d'autocorrélation est définie positive et symétrique. En fait, elle possède une double symétrie, puisqu'elle est aussi symétrique par rapport à la seconde diagonale. Il en résulte un ensemble de propriétés fondamentales. [2].

On considère d'abord les valeurs propres $\lambda_i (0 \leqslant i \leqslant N-1)$ de la matrice d'autocorrélation d'ordre N. L'équation caractéristique :

$$\text{dét}(\lambda I_N - R_N) = 0$$

conduit aux relations :

$$\text{dét}(R_N) = \prod_{i=0}^{N-1} \lambda_i \tag{13.23}$$

$$N . r(0) = \sum_{i=0}^{N-1} \lambda_i = N\sigma_x^2 \tag{13.24}$$

C'est-à-dire que si le déterminant de la matrice est non nul, aucune valeur propre n'est nulle et leur somme est égale à N fois la puissance du signal.

Le caractère défini positif de la matrice R_N entraîne de plus qu'elles sont toutes positives :

$$\lambda_i > 0; \quad 0 \leqslant i \leqslant N-1 \tag{13.25}$$

Pour qu'il en soit ainsi, il faut et il suffit que les déterminants suivants soient tous positifs :

$$r(0)\,;\ \text{dét}\begin{bmatrix} r(0) & r(1) \\ r(1) & r(0) \end{bmatrix};\ \dots;\ \text{dét}\begin{bmatrix} r(0) & r(1) & \dots & r(N-1) \\ r(1) & r(0) & \dots & r(N-2) \\ \vdots & & & \vdots \\ r(N-1) & \dots & & r(0) \end{bmatrix}$$

Les matrices correspondantes sont les matrices d'autocorrélation d'ordre inférieur ou égal à N.

Dans ces conditions, la matrice R_N est diagonalisable et il vient :

$$R_N = M^{-1}.\,\text{diag}(\lambda_i).\,M \tag{13.26}$$

où M est une matrice carrée de dimension N, telle que $M^t = M^{-1}$ et diag(λ_i) la matrice diagonale des valeurs propres. M^t peut être égale à M dans certains cas.
La matrice s'exprime en fonction des vecteurs propres normalisés U_i $(0 \leqslant i \leqslant N-1)$ par :

$$R_N = \sum_{i=0}^{N-1} = \lambda_i U_i U_i^t \tag{13.26 bis}$$

L'analyse des systèmes fait apparaître les puissances successives des matrices R_N et R_N^{-1}. En utilisant, d'une part, le théorème de Cayley-Hamilton selon lequel une matrice vérifie son équation caractéristique et, d'autre part, la formule d'interpolation de Lagrange déjà utilisée au paragraphe V.5, on montre que la puissance d'une matrice s'exprime en fonction des puissances de ses valeurs propres :

$$(R_N)^p = \sum_{i=0}^{N} \lambda_i^p \prod_{\substack{j=0 \\ j \neq i}}^{N-1} \frac{R_N - \lambda_j I_N}{\lambda_i - \lambda_j} \tag{13.27}$$

Pour les grandes valeurs de l'entier p, avec :

$$\lambda_{max} = \max_{0 \leqslant i \leqslant N-1} (\lambda_i)$$

on peut écrire, si ce maximum correspond à la valeur zéro de l'indice :

$$(R_N)^p \simeq \lambda_{max}^p \sum_{j=1}^{N-1} \frac{R_N - \lambda_j I_N}{\lambda_{max} - \lambda_j} \tag{13.28}$$

Par suite pour les grandes valeurs de p, on peut faire l'approximation :

$$(R_N)^p \simeq \lambda_{max}^p.\,K_N \tag{13.29}$$

où K_N est la matrice carrée d'ordre N de la relation (13.28) ; comme la matrice R_N est diagonalisable et satisfait (13.26), K_N s'exprime aussi plus simplement comme le produit de M^{-1} par une matrice déduite de M en annulant toutes les lignes sauf celle qui correspond à l'indice de la plus grande valeur propre.

De même d'après (13.26) on peut écrire dans les mêmes conditions :

$$(R_N^{-1})^p \simeq \lambda_{max}^{-p}.\,K'_N \tag{13.30}$$

avec :

$$\lambda_{\min} = \min_{0 \leqslant i \leqslant N-1} (\lambda_i)$$

En fait, il apparaît dans la suite que ces deux valeurs propres extrêmes, $\lambda_{\min}$ et $\lambda_{\max}$, conditionnent le comportement des systèmes adaptatifs.

L'interprétation physique des valeurs propres de la matrice d'autocorrélation n'apparaît pas aisément à partir de leur définition. Pour éclairer ce point il est intéressant de les rapprocher du spectre du signal $x(n)$.

Le cas où le signal $x(n)$ est périodique et de période N est considéré d'abord. Alors, la suite $r(p)$ elle aussi est périodique, et, de plus, elle est symétrique avec :

$$r(N-i) = r(i)\,; \quad 0 \leqslant i \leqslant N-1$$

Dans ces conditions la matrice R_N est une matrice circulante, dans laquelle chaque ligne se déduit de la précédente par décalage. Si la suite $\Phi_{xx}(n)$ $(0 \leqslant n \leqslant N-1)$ désigne la transformée de Fourier de la suite $r(p)$, la relation suivante est facile à vérifier directement, T_N étant la matrice de la transformée de Fourier d'ordre N :

$$R_N \,.\, T_N = T_N \,.\, \mathrm{diag}\,(\Phi_{xx}(n))$$

d'où :

$$R_N = T_N \,.\, \mathrm{diag}\,(\Phi_{xx}(n)) \,.\, T_N^{-1} \tag{13.31}$$

D'après 13.26, les valeurs propres de la matrice R_N sont dans ce cas les valeurs de la transformée de Fourier discrète de la fonction d'autocorrélation, c'est-à-dire les valeurs de la densité spectrale de puissance du signal. M est la matrice de la transformée en cosinus (par. 3.3.3).

Cette relation est également valable pour un bruit blanc discret puisque le spectre est constant et que, comme la matrice d'autocorrélation est à un facteur près une matrice unité, les valeurs propres sont égales.

Les signaux réels ont généralement une densité spectrale de puissance non constante et leur fonction d'autocorrélation $r(p)$ décroît quand l'indice p croît. Alors, pour N suffisamment grand, les éléments significatifs de la matrice de dimension N se trouvent regroupés au voisinage de la diagonale principale. Dans ces conditions, soit R'_N la matrice d'autocorrélation du signal $x(n)$ supposé périodique et de période N; ses valeurs propres forment un échantillonnage de la densité spectrale de puissance. La différence entre R_N et R'_N tient au fait que R'_N est une matrice circulante et elle apparaît principalement dans le coin supérieur droit et le coin inférieur gauche; ainsi on peut observer que R_N se rapproche davantage d'une matrice diagonale que R'_N; par suite ses valeurs propres sont moins dispersées. En fait, sous certaines conditions généralement réalisées en pratique, on montre que :

$$\min_{0 \leqslant n \leqslant N-1} (\Phi_{xx}(n)) \leqslant \lambda_{\min} \leqslant \lambda_{\max} \leqslant \max_{0 \leqslant n \leqslant N-1} (\Phi_{xx}(n)) \tag{13.32}$$

et pour N suffisamment grand :

$$\lambda_{\min} \simeq \min_{0 \leqslant f \leqslant 1} (\Phi_{xx}(f)); \quad \lambda_{\max} \simeq \max_{0 \leqslant f \leqslant 1} (\Phi_{xx}(f)) \tag{13.33}$$

Finalement, on peut considérer en pratique que les valeurs propres extrêmes de la matrice d'autocorrélation approchent les valeurs extrêmes de la densité spectrale de puissance du signal, quand la dimension de cette matrice est suffisamment grande.

13.4 MODÉLISATION

Les filtres numériques s'appliquent à la modélisation des systèmes, selon le schéma de la figure 13.1. Le signal x(n) est appliqué au système et au modèle et les coefficients sont calculés pour minimiser l'écart entre les sorties du système et du modèle.

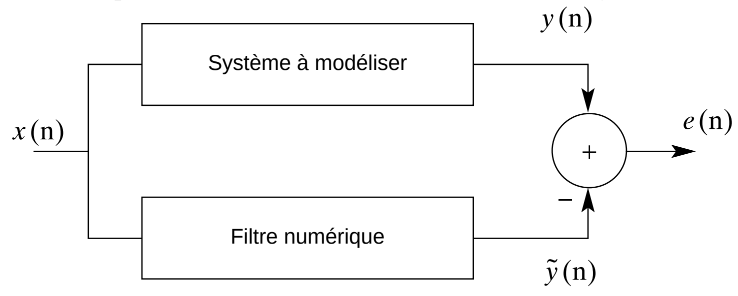


FIG. 13.1. *Modélisation d'un système*

Le type de filtre à utiliser comme modèle dépend de la connaissance a priori du système à modéliser. Cependant, les filtres de type RIF s'imposent en général, pour leur facilité de calcul et de mise en œuvre. La sortie s'écrit alors :

$$\tilde{y}(n) = \sum_{i=0}^{N-1} h_i x(n\text{-}i) = H^t X(n) \tag{13.34}$$

où X(n) est le vecteur des N données les plus récentes, le nombre de coefficients N étant choisi en fonction de la connaissance du modèle. L'erreur de sortie est définie par

$$\mathrm{e}(n) = \mathrm{y}(n) - \tilde{y}(n) \tag{13.35}$$

Le critère retenu pour le calcul des coefficients est généralement la minimisation de l'erreur quadratique moyenne (EQM) :

$$J = E[e^2(n)] \tag{13.36}$$

L'annulation des dérivées de la fonction coût J conduit à la relation $E[e(n)X(n)] = 0$, qui exprime la décorrélation entre la sortie et le vecteur des données d'entrée les plus récentes. Il vient alors :

$$E[y(n)\,X(n)] - E[X(n)\,X^t(n)]\,H = 0 \qquad (13.37)$$

La définition (13.1) de la fonction d'autocorrélation montre que $E[X(n)\,X^t(n)] = R_N$ et les coefficients sont donnés par :

$$H = R_N^{-1} r_{yx} \qquad (13.38)$$

c'est-à-dire que les coefficients du filtre modèle sont obtenus en multipliant l'inverse de la matrice AC par le vecteur d'intercorrélation entre la sortie du système et son entrée, défini par :

$$r_{yx} = E[\,y(n) \begin{bmatrix} x(n) \\ x(n-1) \\ . \\ . \\ . \\ x(n+1)-N \end{bmatrix}] \qquad (13.39)$$

L'erreur quadratique moyenne minimale, E_{min}, est obtenue en combinant (13.36) et (13.38), ce qui fournit les 3 expressions :

$$\begin{aligned} E_{min} &= E[y^2(n)] - H^t R_N H \\ E_{min} &= E[y^2(n)] - H^t r_{yx} \\ E_{min} &= E[y^2(n)] - r_{yx}^t R_N^{-1} r_{yx} \end{aligned} \qquad (13.40)$$

L'égalisation est un cas particulier dans lequel le système à modéliser est l'inverse de celui qui a produit le signal d'entrée $x(n)$. Ainsi, en transmission et en l'absence de bruit, l'égaliseur a pour fonction de transfert l'inverse de celle du canal.

Exemple :
Soit le signal $x(n)$ en sortie d'un canal, lié aux données émises $d(n)$, supposées non corrélées et de puissance unité, par la relation :

$$x(n) = d(n) + 0{,}5\,d(n-1) + 0{,}2\,d(n-2)$$

Les 3 premiers éléments de la fonction AC ont pour valeur :

$$r(0) = 1{,}29 \quad ; \quad r(1) = 0{,}60 \quad ; \quad r(2) = 0{,}20$$

En prenant comme signal de référence $y(n) = d(n)$, on obtient les 3 coefficients d'un égaliseur par :

$$H = \begin{bmatrix} 1{,}29 & 0{,}60 & 0{,}20 \\ 0{,}60 & 1{,}29 & 0{,}60 \\ 0{,}20 & 0{,}60 & 1{,}29 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0{,}9953 \\ -0{,}4971 \\ 0{,}0778 \end{bmatrix}$$

Comme erreur de sortie, on trouve $E_{\min} = 0{,}0047$ et la fonction de transfert T(Z) de l'ensemble canal et égaliseur s'exprime par :

$$T(Z) = 0{,}9953 - 0{,}0015\,Z^{-1} + 0{,}0273\,Z^{-3} - 0{,}0609\,Z^{-3} + 0{,}0156Z^{-4}$$

On peut peut vérifier que le filtre H(Z) constitue une approximation à 3 coefficients de l'inverse du canal, qui est un filtre à réponse impulsionnelle infinie.

13.5 PRÉDICTION LINÉAIRE

La prédiction linéaire est un cas particulier de modélisation dans le lequel la sortie du système à modéliser est le signal lui-même, comme le montre la figure 13.2

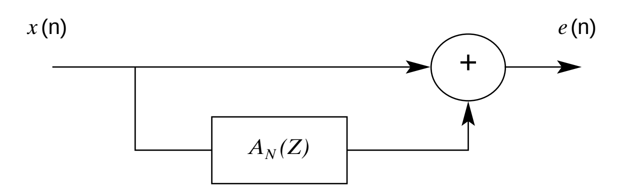


FIG. 13.2. *Principe de la prédiction linéaire*

L'erreur de sortie, ou erreur de prédiction, s'écrit

$$e(n) = x(n) - \sum_{i=1}^{N} a_i x(n-i) \tag{13.41}$$

et les coefficients de prédiction sont donnés par

$$A_N = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} = R_N^{-1} \begin{bmatrix} r(1) \\ \vdots \\ r(N) \end{bmatrix} \tag{13.42}$$

La décorrélation entre l'erreur de prédiction et le signal d'entrée implique les N relations :

$$r(p) = \sum_{i=1}^{N} a_i r(p-i) \quad ; \quad 1 \leq p \leq N \tag{13.43}$$

Pour l'erreur quadratique moyenne minimale (EQMM), il vient :

$$E_{aN} = r(0) - \sum_{i=1}^{N} a_i r(i) \qquad (13.44)$$

Les relations ci-dessus se combinent pour fournir l'équation matricielle de la prédiction linéaire :

$$R_{N+1} \begin{bmatrix} 1 \\ -A_N \end{bmatrix} = \begin{bmatrix} E_{aN} \\ 0 \end{bmatrix} \qquad (13.45)$$

Un signal est dit prédictible si l'erreur de prédiction est nulle, c'est-à-dire s'il obéit à l'équation de récurrence suivante :

$$x(n) = \sum_{i=1}^{N} a_i x(n-i)$$

Le signal est alors constitué de sinusoïdes en nombre inférieur ou égal à N/2 et le filtre de fonction de transfert $H(Z) = 1 - A_N(Z)$ possède des zéros sur le cercle unité , N au maximum. Par exemple, pour N=2 et une sinusoïde à la fréquence f_0, la récurrence devient :

$$x(n) = 2\cos(2\pi f_0)\, x(n-1) - x(n-2)$$

Les zéros de H(Z) sont soit sur le cercle unité, soit à l'intérieur car il est à phase minimale. En effet, s'il existait un zéro en dehors du cercle unité, $|Z_0| > 1$, alors la fonction H'(Z) telle que

$$H'(Z) = \frac{H(Z)}{(1 - Z_0 Z^{-1})(1 -= \overline{Z}_0 Z^{-1})} \left(1 - \frac{Z^{-1}}{Z_0}\right)\left(1 -= \frac{Z^{-1}}{\overline{Z}_0}\right)$$

conduirait à une erreur de prédiction plus faible que H(Z) puisqu'on aurait :

$$|H'(e^{j\omega})| = \frac{1}{|Z_0|^2} |H(e^{j\omega})|$$

comme indiqué au paragraphe 9.6 . D'après l'expression de la puissance du signal de sortie d'un filtre donnée par la relation 4.24, la puissance en sortie de H'(Z) serait inférieure à la puissance en sortie de H(Z), pour un même signal appliqué aux deux filtres.

Le filtre H(Z) étant à phase minimale et inversible, la prédiction linéaire est utilisée pour analyser et modéliser les signaux. En effet, à partir de l'erreur de prédiction on peut remonter au signal en inversant la relation (13.41).

$$x(n) = e(n) - \sum_{i=1}^{N} a_i x(n-i) \qquad (13.46)$$

C'est la modélisation dite auto régressive (AR). En faisant l'hypothèse que l'erreur de prédiction e(n) est un bruit blanc de puissance E_{aN}, une estimation du spectre du signal est donnée par :

$$S_{AR}(f) = \frac{E_{aN}}{\left|1 - \sum_{i=1}^{N} a_i e^{-j2\pi if}\right|^2} \tag{13.47}$$

Si le filtre H(Z) est de type RII, le modèle est dit autorégressif à moyenne adaptée (ARMA) et on peut également en déduire une estimation spectrale.

13.6 STRUCTURES DE PRÉDICTEUR

Les filtres RIF et RII peuvent être réalisés par des structures en treillis, comme indiqué au paragraphe 8.5. L'approche est particulièrement avantageuse en prédiction linéaire. [3]

Les coefficients de prédiction sont données par la relation (13.42). L'inversion de matrice qu'implique cette expression peut être évitée par une procédure itérative.

L'algorithme de Levinson-Durbin fournit une solution du système (13.43) par récurrence, en un ensemble de N étapes. Elle est initialisée en posant comme puissance du signal d'erreur :

$$E_0 = r(0)$$

A l'étape de rang $i\,(1 \leqslant i \leqslant N)$ les calculs suivants sont effectués :

$$k_i = \frac{1}{E_{i-1}}\left[r(i) - \sum_{j=1}^{i-1} a_j^{i-1} r(i-j)\right]; \quad 1 \leqslant i \leqslant N$$

$$a_i^i = k_i$$

$$a_j^i = a_j^{i-1} - k_i a_{i-j}^{i-1}; \quad 1 \leqslant j \leqslant i-1 \tag{13.48}$$

$$E_i = (1 - k_i^2)\,.\,E_{i-1} \tag{13.49}$$

A l'étape de rang N, les N coefficients a_i sont obtenus par :

$$a_i = a_i^N; \quad 1 \leqslant i \leqslant N$$

Le terme E_i correspond à la puissance de l'erreur résiduelle avec un prédicteur d'ordre i. D'autre part les valeurs des coefficients k_i obtenues aux étapes précédentes ne sont pas remises en cause à l'étape de rang i; la procédure est bien séquentielle et le modèle s'affine quand le nombre d'étapes, donc de coefficients

augmente, car avec $|k_i| < 1$ la puissance de l'erreur diminue à chaque étape d'après la relation (13.49).
Les coefficients k_i définissent complètement le filtre et conduisent à un mode de réalisation. En effet, le signal d'erreur à l'étape i est la suite $e_i(n)$ telle que :

$$e_i(n) = x(n) - \sum_{j=1}^{i} a_j^i \, x(n-j)$$

La fonction de transfert du filtre correspondant est maintenant désignée par $A_i(Z)$ qui s'écrit donc :

$$A_i(Z) = 1 - \sum_{j=1}^{i} a_j^i \, Z^{-j}$$

D'après la relation (13.48), il vient :

$$A_i(Z) = A_{i-1}(Z) - k_i Z^{-i} A_{i-1}(Z^{-1}) \qquad (13.50)$$

Avec les éléments du paragraphe 8.5, en posant :

$$B_{i-1}(Z) = Z^{-(i-1)} A_{i-1}(Z^{-1})$$

on obtient :

$$A_i(Z) = A_{i-1}(Z) - k_i Z^{-1} B_{i-1}(Z) \qquad (13.51)$$

A la fonction $B_i(Z)$ correspond la suite $b_i(n)$, pour laquelle on établit la relation :

$$b_i(n) = b_{i-1}(n-1) - k_i \, . \, e_{i-1}(n) \qquad (13.52)$$

Finalement les coefficients k_i conduisent à une structure en treillis conforme à la figure 13.3.

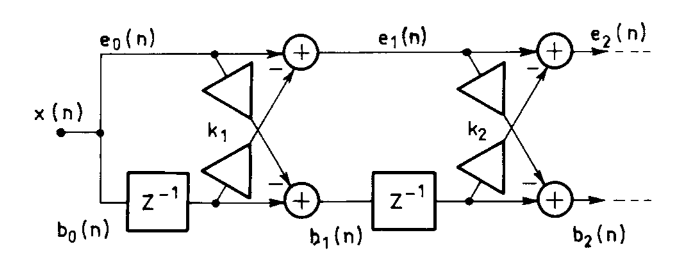


FIG. 13.3. *Filtre de prédiction linéaire en treillis*

La convergence de la procédure est assurée si les relations suivantes sont vérifiées :

$$|k_i| < 1 \, ; \quad 1 \leqslant i \leqslant N \qquad (13.53)$$

Une autre décomposition du filtre de prédiction est fournie par la méthode des paires de raies spectrales, qui consiste à séparer la fonction de transfert en deux parties, associées aux parties paire et impaire des coefficients. En prenant la récurrence (13.50) à l'ordre N+1 et en désignant par $P_N(Z)$ le polynôme obtenu avec $k_{N+1} = 1$, il vient :

$$P_N(Z) = A_N(Z) - Z^{-(N+1)} A_N(Z^{-1}) \qquad (13.54)$$

De même, en désignant par $Q_N(Z)$ le polynôme obtenu avec $k_{N+1} = -1$:

$$Q_N(Z) = A_N(Z) + Z^{-(N+1)} A_N(Z^{-1}) \qquad (13.55)$$

Il s'agit bien d'une décomposition du polynôme $A_N(Z)$ puisque la somme des 2 relations précédentes donne :

$$A_N(Z) = \frac{1}{2}\left[P_N(Z) + Q_N(Z)\right] \qquad (13.56)$$

et $P_N(Z)$ et $Q_N(Z)$ sont des polynômes à coefficients ayant les symétries paire et impaire respectivement. Ils sont à phase linéaire et, comme ils ne peuvent avoir de zéros à l'extérieur du cercle unité en tant que prédicteurs, tous leurs zéros sont sur le cercle unité. De plus, si N est pair, on vérifie les relations : $P_N(1) = 0 = Q_N(-1)$. D'où la factorisation :

$$P_N(z) = (1 - z^{-1}) \sum_{i=1}^{N/2} (1 - 2\cos\theta_i\, z^{-1} + z^{-2}) \qquad (13.57)$$

$$Q_N(z) = (1 + z^{-1}) \sum_{i=1}^{N/2} (1 - 2\cos\omega_i\, z^{-1} + z^{-2})$$

Les deux jeux de paramètres θ_i et ω_i, $1 \le i \le N$, donnent une nouvelle représentation des paramètres de prédiction.

Si $Z_0 = e^{d\omega_0}$ est un zéro du polynôme $A_N(Z)$ sur le cercle unité, c'est aussi un zéro de $P_N(z)$ et $Q_N(z)$. Si ce zéro se déplace vers l'intérieur du cercle unité, les zéros correspondants de $P_N(z)$ et $Q_N(z)$ se déplacent sur le cercle unité, dans des directions opposées à partire de ω_0. En fait une condition nécessaire et suffisante pour que le polynôme $A_N(z)$ soit à phase minimale est que les zéros de $P_N(z)$ et $Q_N(z)$ soient simples et alternent sur le cercle unité.

L'approche ci-dessus conduit à une structure de réalisation du filtre de prédiction donnée à la figure 13.4. Les fonctions de transfert $F(z)$ et $G(z)$ sont les facteurs à phase linéaire des expressions (13.57). La structure se prête à une réalisation en cascade de cellules du second ordre et la propriété de phase minimale globale est contrôlée en vérifiant l'alternance des coefficients des termes en z^{-1}.

Fig. 13.4. *Prédicteur à décomposition polynomiale*

13.7 CONCLUSION

La matrice d'autocorrélation (AC) intervient directement dans la modélisation avec le critère de l' erreur quadratique moyenne minimale (EQMM). Bien qu'elle n'apparaisse pas explicitement dans les algorithmes de filtrage adaptatif les plus courants, ses valeurs propres, en particulier la valeur propre miunimale, conditionnent le fonctionnement du système.

La matrice AC peut même être utilisée directement pour obtenir une analyse spectrale à haute résolution du signal, par la méthode de décomposition harmonique [4]. Dans cette méthode, le signal est modélisé par un ensemble de sinusoides dans du bruit, les fréquences des sinusoides étant associées aux zéros du filtre propre minimal, c'est à dire le filtre qui a pour coefficients les éléments du vecteur propre associé à la valeur propre minimale. Par rapport à la prédiction linéaire, la décomposition harmonique permet d'éviter le biais introduit par le critère des moindres carrés.

La prédiction linéaire permet une analyse des signaux en temps réel et peut fournir un mode de compression simple et efficace.

BIBLIOGRAPHIE

[1] J.Max et collaborateurs, « Méthodes et techniques de traitement du signal », Editions Masson, Paris, 1981.

[2] P.G.Ciarlet, « Introduction à l'analyse numérique matricielle et à l'optimisation », Editions Masson, Paris, 1982.

[3] J.Makhoul, « Linear Prediction : A Tutorial Review », Proceedings of the IEEE, Vol.63, April 1975, pp.561-580.

[4] S.Marcos, « Les méthodes à haute résolution – traitement d'antenne et analyse spectrale », Editions Hermès, Paris, 1998.

EXERCICES

1 Soit le signal $x(n) = \sqrt{2}\sin(n\pi/4)$. Calculer les 3 premiers éléments de la fonction d'autocorrélation. Calculer les valeurs propres et les vecteurs propres de la matrice AC de dimension 3x3. Vérifier les relations 13.26 de décomposition et reconstitution de la matrice.

2 Le prédicteur du second ordre de fonction de transfert $H(Z) = 1 - a_1 Z^{-1} - a_2 Z^{-2}$ est appliqué au signal $x(n) = \sqrt{2}\sin(nw) + b(n)$ où b(n) est un bruit blanc gaussien de puissance σ_b^2. Donner l'expression de la puissance du signal en sortie du prédicteur. Par dérivation, en déduire les expressions des coefficients de prédiction a_1 et a_2. Comment évoluent ces valeurs quand la puissance du bruit augmente.

3 Soit le signal suivant : $x(n)\sin(n\pi/4) + \cos(n\pi/3)$. Calculer les 3 premiers éléments de la fonction d'autocorrélation. En déduire les valeurs des coefficients du prédicteur du second ordre. Placer les zéros du filtre de prédiction dans le plan des Z.

Donner l'expression du signal d'erreur de prédiction et calculer sa puissance.

4 Donner la décomposition polynomiale à symétrie paire et impaire du filtre de prédiction suivant :

$$1 - A_N(Z) = (1 - 1{,}6\,Z^{-1} + 0{,}9\,Z^{-2})(1 - Z^{-1} + Z^{-2})$$

Localiser les zéros des polynômes obtenus. Comparer avec ceux du filtre de départ.

Chapitre 14

Filtrage Adaptatif

Le filtrage adaptatif intervient quand il faut réaliser, simuler ou modéliser un système dont les caractéristiques évoluent dans le temps. Il conduit à la mise en œuvre de filtres à coefficients variables dans le temps. Les variations des coefficients sont définies par un critère d'optimisation et réalisées suivant un algorithme d'adaptation, qui sont déterminés en fonction de l'application. Il existe une grande variété de critères et d'algorithmes possibles [1-4]. Dans le présent chapitre, est considéré le cas simple mais d'une grande importance pratique du critère de minimisation de l'erreur quadratique moyenne associé à l'algorithme du gradient.

14.1 PRINCIPE DU FILTRAGE ADAPTATIF PAR ALGORITHME DU GRADIENT

Le principe du filtrage adaptatif est représenté sur la figure 14.1; il correspond à une opération effectuée sur un signal reçu $x(n)$ pour fournir une sortie dont la différence avec un signal de référence $y(n)$ soit minimisée. Cette minimisation est obtenue en calculant les coefficients du filtre pour chaque nouvel ensemble de données, référence et signal reçu.

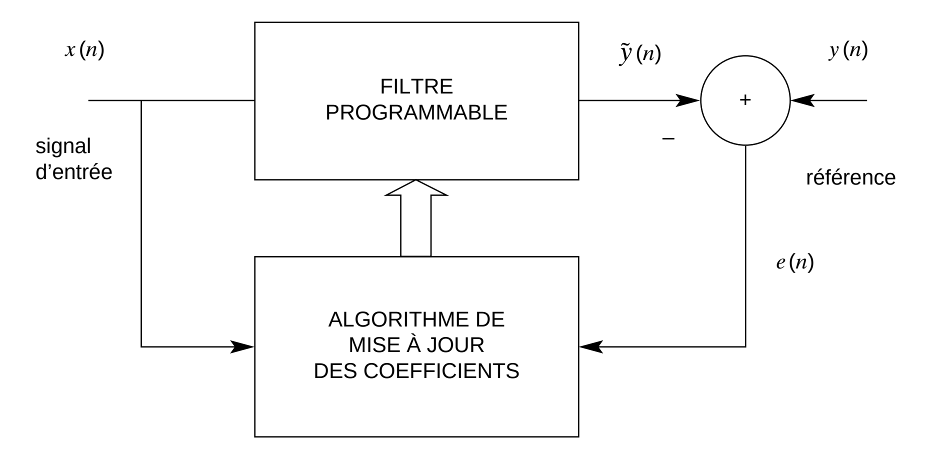


FIG. 14.1. *Principe du filtrage adaptatif*

Ainsi, en supposant qu'à l'indice n, n ensembles de données aient été reçus, les coefficients du filtre adaptatif supposé de type RIF, représentés par le vecteur $H(n)$, qui minimisent la fonction coût quadratique $J(n)$ définie par :

$$J(n) = \sum_{p=1}^{n} [y(p) - H^t(n) X(p)]^2 \tag{14.1}$$

où $X(p)$ est le vecteur colonne d'éléments

$$(x(p), x(p-1), \dots x(p+1-N)),$$

sont donnés, en reprenant les calculs du paragraphe 13.4, par l'équation :

$$H(n) = R_N^{-1}(n) R_{yx}(n) \tag{14.2}$$

$1/2\delta\sigma_x^2$

L'estimation de la matrice d'autocorrélation du signal reçu peut s'exprimer commodément sous la forme :

$$R_N(n) = \sum_{p=1}^{n} X(p) X^t(p) = \sum_{p=1}^{n} \begin{bmatrix} x(p) \\ x(p-1) \\ \vdots \\ x(p+1-N) \end{bmatrix} [x(p), \dots, x(p+1-N)] \tag{14.3}$$

De même, l'estimation du vecteur d'intercorrélation entre référence et entrée s'écrit :

$$r_{yx}(n) = \sum_{p=1}^{n} y(p) X(p) \tag{14.4}$$

Quand le nouvel ensemble de données $\{x(n+1), y(n+1)\}$ devient disponible, le vecteur des coefficients $H(n+1)$ peut être calculé à partir de $H(n)$, par une mise à jour. En effet, d'après les relations (14.3) et (14.4) il vient :

$$R_N(n+1) = R_N(n) + X(n+1)X^t(n+1);$$
$$r_{yx}(n+1) = r_{yx}(n) + X(n+1)y(n+1) \tag{14.5}$$

Et par suite :

$$R_N(n+1)H(n+1) = r_{yx}(n+1) = r_{yx}(n) + X(n+1)y(n+1)$$

soit :

$$R_N(n+1)H(n+1) = R_N(n)H(n) + X(n+1)y(n+1)$$

soit encore :

$$R_N(n+1)H(n+1) = [R_N(n+1) - X(n+1)X^t(n+1)]H(n) + X(n+1)y(n+1)$$

et finalement :

$$H(n+1) = H(n) + R_N^{-1}(n+1)X(n+1)[y(n+1) - H^t(n)X(n+1) \tag{14.6}$$

Il est intéressant de remarquer que la quantité :

$$e(n+1) = y(n+1) - H^t(n)X(n+1) \tag{14.7}$$

représente l'erreur en sortie du système, calculée à l'indice $(n+1)$, avec les coefficients $H(n)$ obtenus à l'indice n ; cette erreur est appelée l'erreur « a priori », alors que le même calcul avec $H(n+1)$ correspond à l'erreur dite « a posteriori ».

Les algorithmes dans lesquels les coefficients sont, à chaque valeur de l'indice, calculés par la récurrence (14.6) sont les algorithmes de moindres carrés.

Des algorithmes simplifiés, mais d'un grand intérêt pratique, sont obtenus en remplaçant la matrice $R_N^{-1}(n)$ par la matrice diagonale δI_N, où δ est un réel que l'on appelle le pas d'adaptation. La mise à jour des coefficients est alors faite par l'équation :

$$H(n+1) = H(n) + \delta X(n+1)e(n+1) \tag{14.8}$$

L'algorithme ainsi obtenu est appelé algorithme du gradient, car la quantité $-X(n+1)e(n+1)$ représente le gradient de la fonction $\frac{1}{2}e^2(n+1)$, c'est-à-dire de la valeur instantanée de l'erreur quadratique. Ainsi la modification des coefficients est faite dans la direction du gradient de l'erreur instantanée, mais avec le signe inverse, ce qui correspond bien à la recherche d'un minimum. Cette procédure est analogue à la méthode dite de plus grande descente en optimisation.

Dans des conditions stationnaires, le vecteur des coefficients converge, en moyenne, vers la solution théorique. En effet la relation (14.8) peut aussi s'écrire, compte tenu de la définition de l'erreur :

$$H(n+1) = [I_N - \delta X(n+1)X^t(n+1)]H(n) + \delta X(n+1)y(n+1) \tag{14.9}$$

En prenant l'espérance des deux membres, puisque :

$$R_N = E[X(n)X^t(n)]; \quad r_{yx} = E[y(n)X(n)] \tag{14.10}$$

où R_N est la matrice d'autocorrélation du signal reçu et r_{yx} le vecteur des N premiers éléments de la fonction d'intercorrélation entre référence et signal reçu, il vient, quand n tend vers l'infini :

$$E[H(\infty)] = H_{opt} = R_N^{-1} r_{yx} \tag{14.11}$$

Ainsi l'algorithme du gradient converge en moyenne vers la solution optimale H_{opt}, d'où la dénomination également de gradient stochastique. Le critère de minimisation correspondant est le critère des moindres carrés moyens.

Une fois la convergence obtenue, les valeurs optimales des coefficients s'expriment par la relation (14.11).

La valeur minimale E_{min} de l'erreur quadratique correspondant à l'ensemble des valeurs optimales des coefficients s'exprime également en fonction des signaux $y(n)$, $x(n)$ et de leur intercorrélation comme indiqué au paragraphe (13.4).

Le schéma du filtre adaptatif obtenu est donné à la figure 14.2. Les variations des coefficients sont calculées par multiplication pour chaque valeur de l'écart $e(n)$ et accumulées.

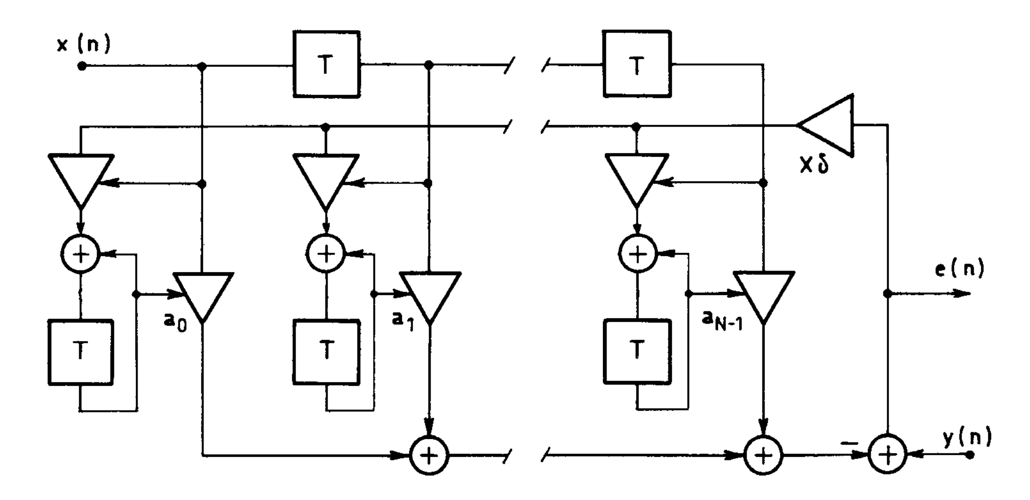


FIG. 14.2. *Filtre RIF adaptatif en structure directe*

Le choix de la valeur δ dans (14.8) résulte d'un compromis entre la rapidité de convergence et la valeur de l'erreur résiduelle quand la convergence est obtenue. Ces deux caractéristiques vont être étudiées successivement, mais il faut d'abord examiner les conditions de convergence.

14.2 CONDITIONS DE CONVERGENCE

Le cas d'un système parfaitement dimensionné et sans bruit de mesure est considéré d'abord. C'est-à-dire qu'après convergence, l'erreur résiduelle est nulle. En désignant le vecteur d'écart des coefficients par $\Delta H(n) = H_{opt} - H(n)$, on peut écrire à l'instant n :

$$\Delta H(n+1) = \Delta H(n) - \delta e(n+1) X(n+1) \tag{14.12}$$

On a également:

$$y(n+1) = H^t_{opt} X(n+1)\,; \quad e(n+1) = \Delta H^t(n)\, X(n+1) \tag{14.13}$$

La norme du vecteur d'écart des coefficients s'écrit :

$$\|\Delta H(n+1)\|_2^2 = \|\Delta H(n)\|_2^2 + \delta^2 e^2(n+1)\, X^t(n+1) X(n+1) - 2\delta e(n+1)\, \Delta^t H(n) X(n+1)$$

et, en utilisant (14.13) :

$$\|\Delta H(n+1)\|_2^2 = \|\Delta H(n)\|_2^2 + \delta e^2(n+1) [\delta X^t(n+1) X(n+1) - 2] \tag{14.14}$$

Une suite monotone décroissante est obtenue et la convergence est garantie si les conditions suivantes sont vérifiées :

$$0 < \delta < \frac{2}{X^t(n+1) X(n+1)} \tag{14.15}$$

On peut en déduire les conditions pour le choix du pas d'adaptation :

$$0 < \delta < \frac{2}{N.\max[x^2(n)]} \tag{14.16}$$

Pour des signaux à facteur de crête élevée et pour des nombres de coefficients N dépassant quelques unités, la borne supérieure peut être trop limitative et conduire à une adaptation lente.

En considérant la moyenne des écarts et en prenant l'espérance des deux termes de la relation (14.14), en supposant l'indépendance entre $e^2(n+1)$ et $X^t(n+1)X(n+1)$, il apparaît que la convergence ne peut pas se produire si les conditions suivantes ne sont pas vérifiées :

$$0 < \delta < \frac{2}{N\sigma_x^2} \tag{14.17}$$

où σ_x^2 désigne la puissance du signal d'entrée.

Les deux bornes supérieures (14.16) et (14.17) sont reliées par le facteur de crête F_c du signal d'entrée et, dans les applications, on peut retenir une valeur

intermédiaire, par exemple (14.17) avec une marge. À noter que dans le cas de signaux gaussiens, la convergence est démontrée pour :

$$0 < \delta < \frac{1}{3}\,\frac{1}{N\sigma_x^2} \tag{14.18}$$

Si l'on introduit un bruit de mesure, le schéma est celui de la figure 14.3.

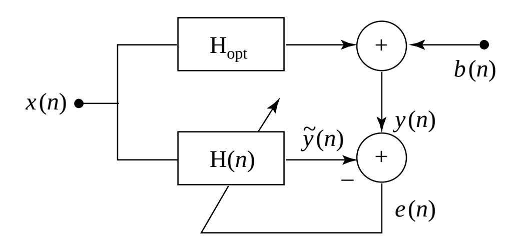


FIG. 14.3. *Filtre adaptatif avec bruit de mesure*

L'erreur de sortie s'écrit :

$$e(n+1) = \Delta H^t(n)\,X(n+1) + b(n+1) \tag{14.14}$$

Le bruit $b(n)$ est de moyenne nulle et a pour puissance σ_b^2. L'écart quadratique des coefficients satisfait la relation de récurrence suivante :

$$\|\Delta H(n+1)\|_2^2 = \|\Delta H(n)\|_2^2 + \delta e^2(n+1)\,[\delta X^t(n+1)\,X(n+1) - 2] \tag{14.20}$$

$$+\, 2\delta b^2(n+1) + 2\delta b(n+1)\,\Delta H^t(n)\,X(n+1)$$

Le bruit $b(n)$ étant non corrélé avec le signal, le dernier terme du membre de droite est nul en moyenne et la condition de convergence s'obtient en prenant l'espérance des deux termes de (14.20) :

$$[\delta N\sigma_x^2 - 2]\,E_R + 2\sigma_b^2 < 0 \tag{14.21}$$

en désignant par E_R l'erreur quadratique moyenne.

Alors, la convergence s'arrête quand l'égalité suivante est vérifiée :

$$E_R = \frac{\sigma_b^2}{1 - \dfrac{\delta}{2} N\sigma_x^2} \tag{14.22}$$

c'est-à-dire qu'après convergence, il subsiste une erreur résiduelle de puissance E_R et la puissance du bruit σ_b^2 correspond à l'erreur quadratique minimale E_{min}, celle qui est atteinte quand les coefficients ont la valeur optimale H_{opt}. Cette erreur résiduelle est analysée dans le cas général au paragraphe 14.4.

La stabilité étant assurée, il est intéressant d'évaluer la rapidité de convergence et de faire apparaître la constante de temps du filtre adaptatif.

14.3 CONSTANTE DE TEMPS

Soit d'abord un filtre à un seul coefficient $h_1(n)$. Dans ces conditions, l'équation d'évolution (14.9) s'écrit :

$$h_1(n+1) = [1 - \delta x^2(n+1)]\, h_1(n) + \delta x(n+1)\, y(n+1) \qquad (14.23)$$

Le coefficient de ce filtre a pour valeur moyenne $b = 1 - \delta\sigma_x^2$. Pour une estimation globale de ses caractéristiques, ce filtre peut être assimilé à une cellule RII du premier ordre à coefficient fixe égal à b. Alors, pour δ petit la relation (6.5) donne la constante de temps :

$$\tau \simeq \frac{1}{\delta\sigma_x^2}$$

Ce résultat peut, sous certaines conditions, se généraliser à un filtre à N coefficients.

Soit $[\alpha(n)]$ le vecteur d'écart défini comme suit :

$$[\alpha(n)] = \mathrm{M}[\mathrm{H}_{opt} - \mathrm{H}(n)] \qquad (14.24)$$

où M est la matrice de rotation qui intervient dans la diagonalisation de $\mathrm{R_N}$ donnée par l'équation (13.26).

En utilisant (14.11), on vérifie que l'équation d'évolution des coefficients (14.9) s'écrit, dans l'espace transformé et en moyenne :

$$\mathrm{E}[\alpha(n+1)] = [\mathrm{I_N} - \delta\, \mathrm{diag}(\lambda_i)]\, \mathrm{E}[\alpha(n)] \qquad (14.25)$$

Ce système correspond à N constantes de temps

$$\tau_i = \frac{1}{\delta\lambda_i} \qquad (14.26)$$

et il montre que c'est la plus petite valeur propre λ_{min} de la matrice d'autocorrélation du signal d'entrée qui détermine le temps de convergence du filtre adaptatif.

Le cas le plus favorable est celui où le signal d'entrée est un bruit blanc car, alors, toutes les valeurs propres sont égales à la puissance du signal et l'on a :

$$\tau_e = \frac{1}{\delta\sigma_x^2} \qquad (14.27)$$

Cette expression donne une estimation de la constante de temps du filtre adaptatif qui est ainsi inversement proportionnelle au pas de variation des coefficients. Les équations d'évolution des coefficients se simplifient dans ce cas, de même

que l'évolution de l'erreur quadratique moyenne $E(n)$. Pour cette dernière, en supposant nulles les valeurs initiales des coefficients, il vient :

$$E(n) = (1 - \delta \quad \sigma_x^2)^{2n}\, \sigma_y^2 \tag{14.28}$$

où σ_y^2 désigne la puissance du signal de référence.
Une illustration est donnée à la figure 14.4.

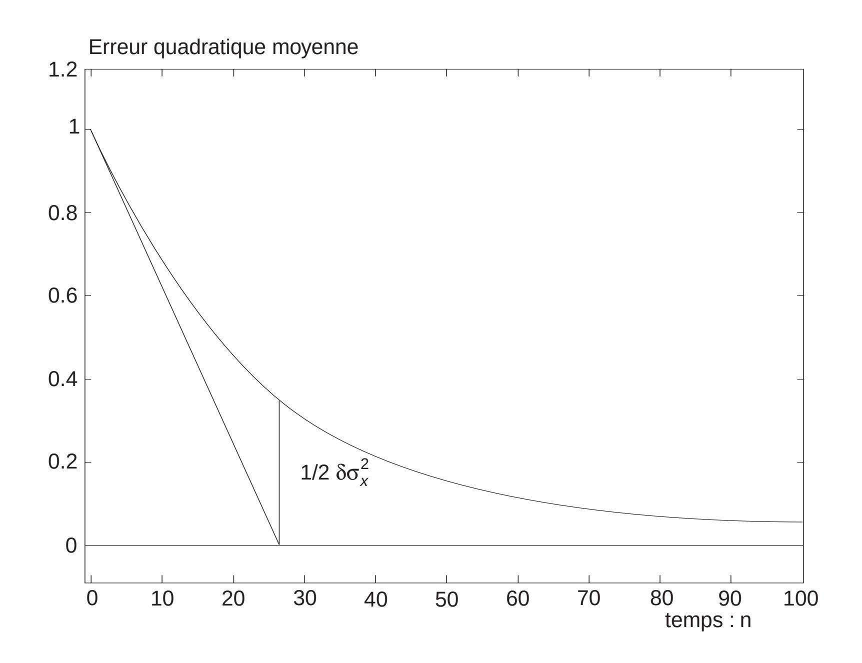


Fig. 14.4. *Évolution de l'EQM pour un bruit blanc en entrée*

Dans les applications, on recherche souvent la plus grande rapidité d'adaptation et l'on souhaite donc donner au pas d'adaptation la plus grande valeur possible.

La condition (14.17) donne une valeur maximale pour le pas d'adaptation qui, avec les hypothèses faites, correspond à une limite de convergence, c'est-à-dire que si δ dépasse cette limite, l'erreur quadratique s'accroît, en moyenne.

Une illustration géométrique simple, basée sur une représentation de la surface d'erreur supposée symétrique, montre que la plus grande rapidité d'adaptation est obtenue pour la moitié de cette limite, soit $\delta = \dfrac{1}{N\sigma_x^2}$. Ce résultat se démontre analytiquement à partir de l'approche donnée au paragraphe suivant. Dans ces conditions la constante de temps satisfait l'inégalité :

$$\tau_e \geqslant N \tag{14.29}$$

Pour compléter l'étude du filtre adaptatif, il faut encore évaluer l'erreur résiduelle après convergence, dans le cas général, c'est-à-dire avec dimensionnement imparfait et bruit de mesure.

14.4 ERREUR RÉSIDUELLE

Après une phase transitoire correspondant à la convergence, les coefficients du filtre adaptatif évoluent en permanence autour de leur valeur optimale, car le pas de variation δ reste constant, ce qui d'ailleurs est la condition d'adaptation permanente du système. Il en résulte que l'erreur résiduelle E_R, définie comme la limite de l'espérance de l'erreur quadratique $E(n)$ quand n tend vers l'infini, reste supérieure à la valeur minimale E_{min}.

L'erreur résiduelle E_R est évaluée en considérant l'évolution du vecteur $[\alpha(n)]$ défini par l'équation (14.24) :

$$[\alpha(n+1)] = [\alpha(n)] - \delta M X(n+1)\, e(n+1) \qquad (14.30)$$

Pour estimer la valeur des carrés des éléments du vecteur $[\alpha(n)]$, il est commode de considérer la matrice $[\alpha(n)]\,[\alpha(n)]^t$ dont la diagonale est constituée des éléments cherchés.

Il vient :

$$[\alpha(n+1)]\,[\alpha(n+1)]^t = [\alpha(n)]\,[\alpha(n)]^t - 2\delta M X(n+1)\, e(n+1)\,[\alpha(n)]^t + \delta^2 e^2(n+1)\, M X(n+1)\, X^t(n+1)\, M^t \qquad (14.31)$$

L'erreur $e(n+1)$ s'exprime en fonction de $[\alpha(n)]$ par :

$$e(n+1) = y(n+1) - H^t_{opt} X(n+1) + X^t(n+1)\, M^t[\alpha(n)]$$

En fait l'évolution du système est commandée par ce couple d'équations. Afin d'obtenir des résultats utiles, il est nécessaire de faire des hypothèses simplificatrices.

Les variables suivantes sont supposées indépendantes :

- l'erreur pour les valeurs optimales des coefficients;
- le vecteur des données : $X(n+1)$;
- l'écart des coefficients par rapport à l'optimum : $H(n) - H_{opt}$.

Ces hypothèses ont pour conséquence :

$$E[[y(n+1) - H^t_{opt} X(n+1)]\, X^t(n+1)\, M^t[\alpha(n)]] = 0 \qquad (14.32)$$

En prenant l'espérance des deux membres de l'équation (14.31) il vient :

$$E\{[\alpha(n+1)]\,[\alpha(n+1)]^t\} = [I_N - 2\delta\ \mathrm{diag}\,(\lambda_i)]\, E\{[\alpha(n)]\,[\alpha(n)]^t\} + \delta^2 E(n)\ \mathrm{diag}(\lambda_i) \qquad (14.33)$$

Et quand n tend vers l'infini, après la phase transitoire :

$$E\{[\alpha(\infty)]\,[\alpha(\infty)^t\} = \frac{\delta}{2}\,E(\infty)\,I_N \tag{14.34}$$

D'après la définition (14.24) du vecteur $[\alpha(n)]$ il vient aussi :

$$E\{[H_{opt} - H(\infty)]\,[H_{opt} - H(\infty)]^t\} = \frac{\delta}{2}\,E(\infty)\,I_N \tag{14.35}$$

Donc, après convergence, les écarts des coefficients sont indépendants et ont même variance.

Il faut maintenant appliquer ces résultats au calcul de la puissance de l'erreur résiduelle.

Pour un écart $\Delta H(n)$ des coefficients, l'erreur résiduelle correspondante s'écrit :

$$E(n) = E_{min} + \Delta H^t(n)\,R_N\,\Delta H(n) \tag{14.36}$$

ou encore, avec (14.24) :

$$E(n) = E_{min} + [\alpha(n)]^t\,\mathrm{diag}\,(\lambda_i)\,[\alpha(n)] \tag{14.37}$$

En effectuant les produits, il vient

$$E(n) = E_{min} + \sum_{i=0}^{N-1} \lambda_i \alpha_i^2(n) \tag{14.38}$$

Comme les écarts des coefficients ont même variance, on peut faire une mise en facteurs et, en utilisant l'expression (14.35), il apparaît que l'erreur résiduelle à l'infini E_R est donnée par :

$$E_R = \frac{E_{min}}{1 - \frac{\delta}{2} N\sigma_x^2} \tag{14.39}$$

A noter que l'on retrouve ainsi la condition de stabilité (14.17).

En pratique, compte tenu de la marge généralement prise sur le pas d'adaptation δ, l'approximation suivante peut être faite :

$$E_R \simeq E_{min}\left(1 + \frac{\delta}{2} N\sigma_x^2\right) \tag{14.40}$$

En fonction de la constante de temps, avec la relation (14.27), il vient :

$$E_R \simeq E_{min}\left(1 + \frac{NT}{2\tau_e}\right) \tag{14.41}$$

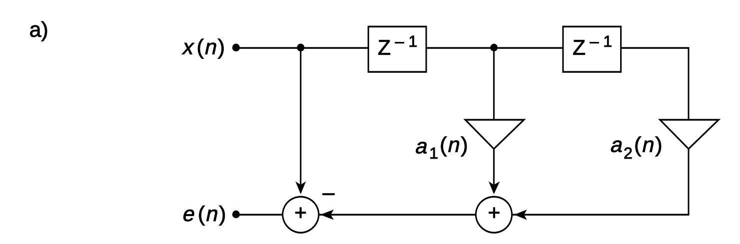
Le compromis entre constante de temps et écart résiduel apparaît ainsi clairement ; T est la période d'échantillonnage considérée jusqu'ici comme égale à l'unité.

L'accroissement d'erreur résiduelle due au pas d'adaptation δ correspond en fait à un bruit de gradient.

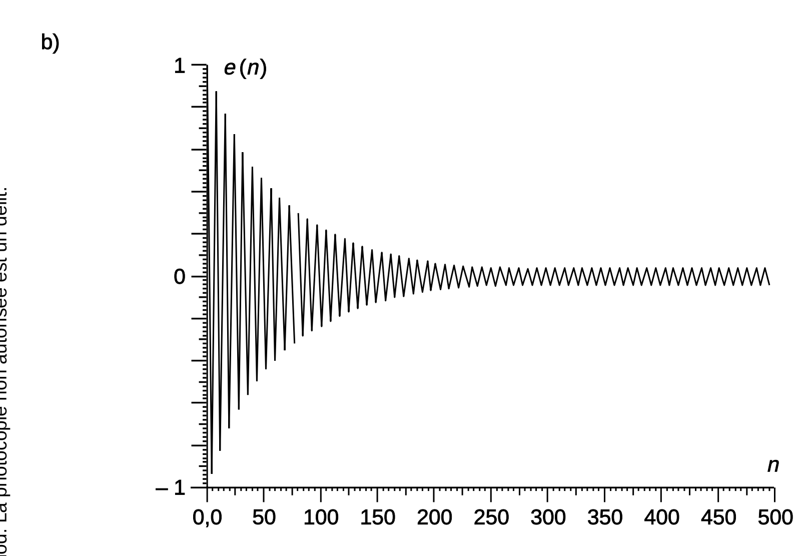
Le fonctionnement d'un filtre adaptatif peut être illustré par le prédicteur d'ordre 2 défini par les équations suivantes :

$$\begin{aligned} e(n+1) &= x(n+1) - a_1(n)x(n) - a_2(n)x(n-1) \\ a_1(n+1) &= a_1(n) + \delta x(n)e(n+1) \\ a_2(n+1) &= a_2(n) + \delta x(n-1)e(n+1) \end{aligned} \tag{14.42}$$

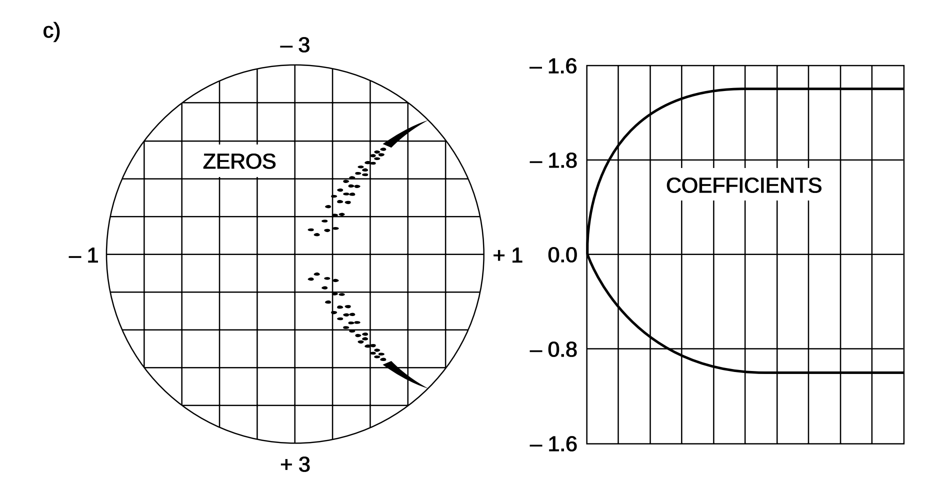
En prenant comme signal $x(n) = \sin n\,\dfrac{\pi}{4}$, avec des valeurs nulles pour les coefficients à l'origine, l'évolution de l'erreur et celle des coefficients sont données par la figure 14.5. Les valeurs optimales des coefficients correspondent à un filtre RIF ayant un zéro sur le cercle unité à la fréquence $\dfrac{f_e}{8}$, pour annuler le signal d'entrée.



a) Schéma du prédicteur du second ordre



b) Erreur en sortie



c) Évolution des coefficients

FIG. 14.5. *Filtre de prédiction du second ordre*

14.5 PARAMÈTRES DE COMPLEXITÉ

Les paramètres de complexité des filtres adaptatifs sont les mêmes que ceux des filtres à coefficients fixes, c'est-à-dire principalement la cadence des multiplications, le nombre de bits des coefficients et des mémoires internes.

Les limitations du nombre de bits des coefficients et des données internes contribuent à augmenter l'erreur résiduelle qui prend la valeur E_{RT}. Les spécifications sont généralement données par un gain minimum du système, c'est-à-dire que le rapport de la puissance du signal de référence σ_y^2 à l'erreur résiduelle totale E_{RT} doit excéder une valeur imposée G^2 :

$$\frac{\sigma_y^2}{E_{RT}} \geqslant G^2 \tag{14.43}$$

La constante de temps τ_e, si elle est imposée, doit être compatible avec le gain minimum du système pour que le filtre soit réalisable avec la technique du gradient.

La donnée des paramètres du filtre adaptatif G et τ_e permet de calculer le nombre de bits des coefficients et des données internes pour chaque structure, l'ordre du filtre étant choisi suffisamment grand pour que l'écart minimal E_{min} soit lui-même suffisamment faible et permette de satisfaire (14.43).

Dans le cas du filtre RIF réalisé en structure directe comme sur la figure 14.2, on effectue généralement les arrondis à la sortie des multiplieurs et le schéma devient celui de la figure 14.6.

FIG. 14.6. *Filtre adaptatif RIF avec limitation des coefficients et données internes*

Le bruit associé à l'arrondi des données internes avec le pas q_2 correspond à l'addition de la puissance $N \frac{q_2^2}{12}$ à l'erreur minimale E_{min}.

Si les coefficients sont quantifiés avec le pas q_1, les erreurs d'arrondi ainsi produites correspondent à un vecteur qu'il faut introduire dans l'équation d'évolution des coefficients. En admettant que les erreurs d'arrondi sont indépendantes des autres signaux et entre elles, il en résulte dans (14.33) un terme supplémentaire égal à $\frac{q_1^2}{12} I_N$ et l'équation (14.34) devient :

$$E\{[\alpha(\infty)]\,[\alpha(\infty)]^t\} = \frac{\delta}{2}\, E(\infty)\, I_N + \frac{1}{2\delta}\,\frac{q_1^2}{12}\,\text{diag}\left(\frac{1}{\lambda_i}\right) \tag{14.44}$$

Toutes opérations faites, l'erreur résiduelle totale E_{RT} s'écrit :

$$E_{RT} = \frac{1}{1 - \frac{\delta}{2} N\sigma_x^2}\left[E_{min} + N\,\frac{q_2^2}{12} + \frac{N}{2\delta}\,\frac{q_1^2}{12}\right] \tag{14.45}$$

Et pour des valeurs petites des pas d'adaptation et de quantification, on peut faire l'approximation :

$$E_{RT} \simeq E_{min}\left(1 + \frac{N.\delta.\sigma_x^2}{2}\right) + \frac{N}{2\delta}\,\frac{q_1^2}{12} + N\,\frac{q_2^2}{12} \tag{14.46}$$

Les valeurs relatives des 4 termes qui interviennent dans cette expression peuvent être choisies en fonction de chaque application. Une option courante consiste à considérer que E_{min} est le plus important et que l'erreur résiduelle supplémen-

taire due au pas d'adaptation δ est égale au bruit introduit par les arrondis internes, c'est-à-dire :

$$\frac{1}{2} \cdot \mathrm{E}_{\min} \cdot \frac{\mathrm{N} \cdot \delta \cdot \sigma_x^2}{2} = \frac{\mathrm{N}}{2\delta} \frac{q_1^2}{12} = \mathrm{N} \frac{q_2^2}{12} \qquad (14.47)$$

Si b_c est le nombre de bits des coefficients et $h_{\max}$ l'amplitude du plus grand coefficient, on a :

$$q_1 = h_{\max} \cdot 2^{1-b_c}$$

Dans ces conditions :

$$2^{2b_c} = \frac{2}{3} \frac{h_{\max}^2}{\delta^2 \cdot \mathrm{E}_{\min} \cdot \sigma_x^2}$$

Avec l'hypothèse que $\mathrm{E}_{\min}$ est le terme prépondérant dans (14.46), c'est-à-dire que l'on a :

$$\mathrm{G}^2 \cdot \mathrm{E}_{\min} \simeq \sigma_y^2$$

et en introduisant la constante de temps, il vient approximativement :

$$b_c \simeq \log 2(\tau_e) + \log 2(\mathrm{G}) + \log 2\left(h_{\max} \cdot \frac{\sigma_x}{\sigma_y}\right) \qquad (14.48)$$

Le terme $\left(h_{\max} \cdot \frac{\sigma_x}{\sigma_y}\right)$ dépend du gain du système, du signal et de l'ordre du filtre, il faut le déterminer pour chaque application.

L'expression (14.48) montre que le nombre de bits des coefficients est directement lié à la constante de temps et au gain du système.

De la même manière les expressions (14.47) permettent de déterminer le nombre de bits des données internes b_i en posant :

$$q_2 = \mathrm{Max}\{|x(n)|, |y(n)|\} \cdot 2^{1-b_i}$$

Avec l'hypothèse $\sigma_x^2 \geqslant \sigma_y^2$, qui correspond notamment au cas de la prédiction linéaire et de l'annulation d'écho, en prenant la valeur 4 pour facteur de crête du signal $x(n)$ comme dans le cas gaussien, il vient :

$$q_2 = 4 \cdot \sigma_x \cdot 2^{1-b_i}$$

et ensuite :

$$2^{2b_i} = 2^4 \cdot \frac{4}{3} \frac{1}{\mathrm{E}_{\min} \cdot \delta}$$

En introduisant les paramètres du système, on obtient approximativement :

$$b_i \simeq 2 + \log 2\left(\frac{\sigma_x}{\sigma_y}\right) + \log 2(\mathrm{G}) + \frac{1}{2} \log 2(\tau_e) \qquad (14.49)$$

Les expressions (14.48) et (14.49) permettent de guider les concepteurs de systèmes dans les options de réalisation.

14.6 ALGORITHMES NORMALISÉS ET ALGORITHMES DU SIGNE

La constante de temps d'un filtre adaptatif et son erreur résiduelle sont liées à la puissance du signal d'entrée $x(n)$. Quand cette puissance peut varier dans des proportions importantes, on peut modifier l'adaptation comme suit :

$$H(n+1) = H(n) + \frac{\delta}{X^t(n+1)X(n+1)} X(n+1)e(n+1) \quad (14.50)$$

C'est un algorithme dit normalisé. On peut vérifier qu'il conduit à une erreur *a posteriori* nulle si $\delta = 1$.

En pratique, plutôt que calculer le produit scalaire $X^t(n+1)X(n+1)$ on peut faire une estimation récursive de la puissance, ce qui conduit à :

$$\begin{aligned} P_x(n+1) &= (1-\varepsilon)P_x(n) + \varepsilon x^2(n+1) \\ H(n+1) &= H(n) + \frac{\delta X(n+1)e(n+1)}{P_x(n+1)} \end{aligned} \quad (14.51)$$

Le paramètre ε de l'estimation récursive est choisi en fonction des variations de la puissance du signal. Il doit au moins être de l'ordre de l'inverse du nombre de coefficients N du filtre.

Dans certaines applications, il est important de minimiser les opérations et on utilise alors des algorithmes simplifiés dans lesquels les variations des coefficients sont fonction du signe des termes $e(n)$ ou $x(n)$, ou encore des produits $e(n).x(n-i)$; ce sont les algorithmes du signe. La réduction de complexité ainsi obtenue se paie par une dégradation de certaines performances du système [2].

Soit par exemple l'algorithme d'adaptation suivant pour les coefficients :

$$h_i(n+i) = h_i(n) + \Delta . e(n+1) . \text{signe}[x(n+1-i)] \quad (14.52)$$

Pour x non nul, on a :

$$\text{signe}[x] = \frac{x}{|x|}$$

Si la loi de probabilité de l'amplitude du signal $x(n)$ est symétrique, en toute première approximation on peut remplacer $|x(n)|$ par la valeur efficace σ_x. Il en résulte que les variations données par l'expression (14.52) sont comparables à celles données par (14.8) avec :

$$\delta \simeq \frac{\Delta}{\sigma_x}$$

Poursuivant dans cette voie, on peut réduire les variations des coefficients à une valeur constante en prenant le produit des signes :

$$h_i(n+1) = h_i(n) + \Delta \,.\, \text{signe}\,[e(n+1)]\,.\,\text{signe}\,[x(n+1-i)] \quad (14.53)$$

Les variations sont alors comparables à celles données par (14.8) avec :

$$\delta \simeq \frac{\Delta}{\sigma_e \sigma_x} \quad (14.54)$$

A partir de valeurs nulles des coefficients, dans la phase de convergence du filtre, on peut considérer que $\sigma_e \simeq \sigma_y$ et la constante de temps τ_s pour l'algorithme du signe considéré peut s'exprimer par :

$$\tau_s \simeq \frac{1}{\Delta}\,\frac{\sigma_y}{\sigma_x} \quad (14.55)$$

Après convergence on peut prendre $\sigma_e^2 \simeq E_{min}$; si le pas de variation est suffisamment petit et pour l'erreur résiduelle E_{RS} dans l'algorithme du signe, il vient :

$$E_{RS} \simeq E_{min}\left(1 + \frac{N\Delta}{2}\,.\,\frac{\sigma_x}{\sqrt{E_{min}}}\right) \quad (14.56)$$

L'erreur résiduelle se trouve ainsi augmentée par rapport à l'algorithme du gradient, ce qui conduit à prendre pour Δ des valeurs faibles. Il faut également noter que la condition de stabilité (14.17) se traduit, en considérant la relation (14.54), par l'inégalité suivante :

$$\Delta \leqslant \frac{2}{N}\,\frac{\sqrt{E_{min}}}{\sigma_x} \quad (14.57)$$

qui fait apparaître une limite inférieure de l'erreur résiduelle et peut conduire à de très petites valeurs de Δ. D'ailleurs, en pratique, on modifie généralement l'équation (14.52) pour effectuer les variations comme suit :

$$h_i(n+1) = (1-\varepsilon)\,.\,h_i(n) + \Delta\,\text{signe}\,[e(n+1)\,.\,x(n+1-i)] \quad (14.58)$$

La constante ε, positive et faible, introduit une fonction de rappel à zéro des coefficients en l'absence de signal.

De plus, dans ces conditions les coefficients sont bornés par :

$$|h_i(n)| \leqslant \frac{\Delta}{\varepsilon}\,; \quad 0 \leqslant i \leqslant N-1 \quad (14.59)$$

Cependant cette modification contribue à augmenter l'erreur résiduelle. En effet il en résulte un biais sur l'estimation des coefficients, car au lieu de (14.11) on a :

$$E\,[H(\infty)] = \left[\frac{\varepsilon}{\Delta}\,I + R_N\right]^{-1}.\,E\,[y(n)\,.\,X(n)] \quad (14.60)$$

L'augmentation correspondante de l'erreur résiduelle se calcule alors par une expression semblable à (14.36).

Le choix des constantes ε et Δ est à faire en fonction des performances à atteindre dans chaque cas. Un choix raisonnable consiste à prendre ε plus petit que Δ, par exemple d'un ordre de grandeur.

Les filtres adaptatifs étudiés dans les paragraphes ci-dessus sont de type RIF en structure directe. C'est une approche simple et robuste, très utilisée. En fait, comme pour les filtres à coefficients fixes on peut faire appel à d'autres structures.

14.7 FILTRAGE RIF ADAPTATIF EN STRUCTURE CASCADE

Dans certains problèmes de modélisation ou d'automatique, il est important de connaître les racines de la fonction de transfert en Z du filtre adaptatif.

Il est alors commode de faire appel à une réalisation par mise en cascade de L cellules du second ordre $H_i(Z)$, $1 \leqslant i \leqslant L$, telles que :

$$H_i(Z) = 1 + a_1^i Z^{-1} + a_2^i Z^{-2}$$

En effet d'après les résultats du chapitre VI, si les zéros de cette cellule Z_1^i et Z_2^i sont complexes, on a :

$$Z_2^i = \overline{Z_1^i}; \quad a_2^i = |Z_1^i|^2; \quad a_1^i = -2Re(Z_1^i)$$

Soit un filtre adaptatif dont la fonction de transfert H(Z) s'écrit :

$$H(Z) = \prod_{i=1}^{L} (1 + a_1^i Z^{-1} + a_2^i Z^{-2}) \tag{14.60}$$

A partir d'un ensemble donné de valeurs des coefficients il faut appliquer des variations proportionnelles au gradient de la fonction d'écart E(A), pour minimiser l'erreur quadratique moyenne. Compte tenu de la relation de définition de E(A), il vient :

$$\frac{\partial E}{\partial a_k^i} = -\frac{2}{N_0} \prod_{n=0}^{N_0-1} [y(n) - \tilde{y}(n)] . \frac{\partial \tilde{y}(n)}{\partial a_k^i} ; \quad \begin{array}{l} k = 1,2 \\ 1 \leqslant i \leqslant L \end{array} \tag{14.61}$$

Pour calculer le terme g_k^i tel que :

$$g_k^i(n) = \frac{\partial \tilde{y}(n)}{\partial a_k^i}$$

on peut faire appel à l'expression de $\tilde{y}(n)$ obtenue par transformation en Z inverse à partir de la transformée X(Z) de la suite $x(n)$. Il vient :

$$\tilde{y}(n) = \frac{1}{2\pi j} \int_{\Gamma} Z^{n-1} . \sum_{i=1}^{L} (1 + a_1^i Z^{-1} + a_2^i Z^{-2}) . X(Z) . dZ$$

où Γ est un contour d'intégration convenable. D'où :

$$\frac{\partial \tilde{y}(n)}{\partial a_k^i} = \frac{1}{2\pi j} \int_\Gamma Z^{n-1}.Z^{-k}. \prod_{\substack{l=1 \\ l \neq i}}^{L} (1 + a_1^l Z^{-1} + a_2^l Z^{-2}).X(Z).dZ$$

ou encore :

$$\frac{\partial \tilde{y}(n)}{\partial a_k^i} = \frac{1}{2\pi j} \int_\Gamma Z^{n-1}.Z^{-k}. \frac{H(Z)}{1 + a_1^i Z^{-1} + a_2^i Z^{-2}} .X(Z).dZ \qquad (14.62)$$

c'est-à-dire que pour former le terme $g_k^i(n)$, il suffit d'appliquer la suite $\tilde{y}(n)$ à une cellule récursive dont la fonction de transfert est l'inverse de celle de la cellule de rang i; c'est une cellule avec les mêmes coefficients, mais de signe opposé. Le schéma correspondant est donné par la figure 14.7. Les variations des coefficients sont calculées par les expressions suivantes :

$$da_k^i(n) = \delta.g_k^i(n).[y(n) - \tilde{y}(n)] : \quad \begin{array}{l} k = 1,2 \\ 1 \leqslant i \leqslant L \end{array} \qquad (14.63)$$

Le filtre ainsi obtenu est plus compliqué que celui du paragraphe précédent, mais il fournit très simplement la position des zéros du filtre, qui, en raison de la présence d'une partie récursive doivent être à l'intérieur du cercle unité dans le plan des Z pour assurer la stabilité du système.

Les techniques élaborées pour les filtres RIF s'étendent aux filtres RII.

FIG. 14.7. *Filtre RIF adaptatif en structure cascade*

14.8 FILTRAGE ADAPTATIF RII

Les coefficients d'un filtre RII peuvent être calculés pour des spécifications dans le temps en utilisant comme au paragraphe 7.3, la technique de minimisation de l'erreur quadratique moyenne dans une procédure itérative. Des algorithmes pour l'adaptation des coefficients aux évolutions d'un système dans le temps s'en déduisent, comme pour la structure RIF.

Un système linéaire peut être modélisé par un filtre RII purement récursif de fonction de transfert en Z, G (Z) telle que :

$$G(Z) = a_0 \cdot \frac{1}{1 + \sum_{k=1}^{K} b_k Z^{-k}} \tag{14.64}$$

Dans ce cas, le modèle est dit autorégressif (AR). Cette approche très intéressante et commode, est encore convenable si la meilleure représentation du système correspond à une fonction de transfert en Z, H (Z) quotient de deux polynômes et telle que :

$$H(Z) = \frac{N(Z)}{D(Z)}$$

où N (Z) a tous ses zéros à l'intérieur du cercle unité et est ainsi à phase minimale. En effet, dans ce cas on peut écrire, pour un entier M convenable :

$$\frac{1}{N(Z)} \simeq 1 + \sum_{i=1}^{M} c_i Z^{-i}$$

Alors il suffit de prendre pour le degré K du dénominateur de la fonction G (Z) une valeur suffisante pour représenter H (Z). La présence de zéros intérieurs au cercle unité dans le système, conduit à une augmentation du nombre de pôles du modèle [1].

Le filtre RII général correspond à un modèle dit autorégressif à moyenne adaptée (ARMA). C'est l'approche la plus générale pour modéliser un système linéaire. La sortie du filtre RII dont il faut calculer les coefficients pour approcher une suite $y(n)$, sur un ensemble de N_0 indices, s'écrit :

$$\tilde{y}(n) = \sum_{l=0}^{L} a_l x(n-l) - \sum_{k=1}^{K} b_k \tilde{y}(n-k) \tag{14.65}$$

La fonction d'erreur E (A, B) s'exprime par :

$$E(A, B) = \frac{1}{N_0} \sum_{n=0}^{N_0-1} [y(n) - \tilde{y}(n)]^2 \tag{14.66}$$

A partir d'un ensemble de valeurs de coefficients, pour minimiser la fonction d'écart suivant l'algorithme du gradient, il faut donner aux coefficients des variations proportionnelles au gradient de la fonction E (A, B) et de signe opposé.

La présence d'une partie récursive introduit des complications. En effet, le calcul du gradient conduit aux expressions :

$$\frac{\partial E}{\partial a_l} = -\frac{2}{N_0} \sum_{n=0}^{N_0-1} [y(n) - \tilde{y}(n)] \frac{\partial \tilde{y}(n)}{\partial a_l} ; \quad 0 \leqslant l \leqslant L$$

$$\frac{\partial E}{\partial b_k} = -\frac{2}{N_0} \sum_{n=0}^{N_0-1} [y(n) - \tilde{y}(n)] . \frac{\partial \tilde{y}(n)}{\partial b_k} ; \quad 1 \leqslant k \leqslant K$$

avec

$$\frac{\partial \tilde{y}(n)}{\partial a_l} = x(n-l) - \sum_{k=1}^{K} b_k . \frac{\partial \tilde{y}(n-k)}{\partial a_l} \tag{14.67}$$

$$\frac{\partial \tilde{y}(n)}{\partial b_k} = -\tilde{y}(n-k) - \sum_{k=1}^{K} b_k . \frac{\partial \tilde{y}(n-k)}{\partial b_k} \tag{14.68}$$

Pour faire apparaître le mode de réalisation des relations (14.67) et (14.68), on pose :

$$H(Z) = \frac{\sum_{l=0}^{L} a_l Z^{-l}}{1 + \sum_{k=1}^{K} b_k Z^{-k}} = \frac{N(Z)}{D(Z)}$$

Il vient :

$$\tilde{y}(n) = \frac{1}{2\pi j} \int_{\Gamma} Z^{n-1} . H(Z) . X(Z) . dZ$$

et par suite :

$$\frac{\partial \tilde{y}(n)}{\partial a_l} = \frac{1}{2\pi j} \int_{\Gamma} Z^{n-1} . Z^{-l} . \frac{1}{D(Z)} . X(Z) . dZ \tag{14.69}$$

$$\frac{\partial \tilde{y}(n)}{\partial b_k} = \frac{1}{2\pi j} \int_{\Gamma} Z^{n-1} . Z^{-k} . \frac{(-1)}{D(Z)} . H(Z) . X(Z) . dZ \tag{14.70}$$

Le gradient est ainsi calculé à partir de la suite obtenue en appliquant les suites $x(n)$ et $\tilde{y}(n)$ aux circuits correspondant à la fonction de transfert $\frac{1}{D(Z)}$.

Une approche simplifiée conduit à ignorer les seconds termes dans (14.67) et (14.68) ce qui correspond à la réalisation la plus simple :

$$da_l(n) = \delta . [y(n) - \tilde{y}(n)] . x(n-l) ; \quad 0 \leqslant l \leqslant L \tag{14.71}$$

$$db_k(n) = -\delta . [y(n) - \tilde{y}(n)] . \tilde{y}(n-k) ; \quad 1 \leqslant k \leqslant K$$

Le schéma correspondant est donné par la figure 14.8.

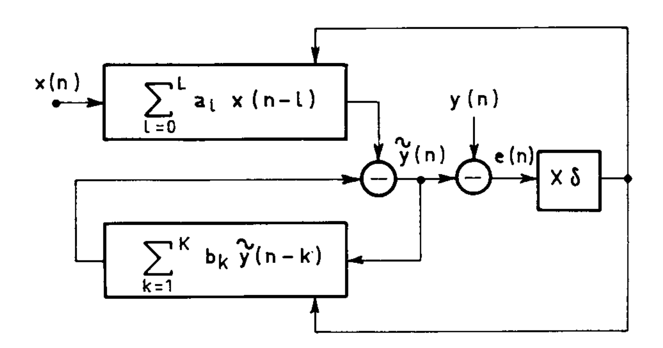


FIG. 14.8. *Filtre RII adaptatif en structure directe*

Pour chaque valeur de l'indice n, les coefficients a_l et b_k sont incrémentés d'une quantité proportionnelle au produit de l'erreur

$e(n) = y(n) - \tilde{y}(n)$ par $x(n-l)$ et $\tilde{y}(n-k)$ respectivement.

L'étude de la stabilité et des paramètres de ce type de filtre adaptatif est plus délicate que pour le filtre de type RIF [5].

Cependant la stabilité d'un filtre RII est facile à contrôler quand il est réalisé par une cascade de cellules du second ordre, ce qui de plus offre les avantages indiqués dans un précédent chapitre pour la réalisation.

Soit un filtre réalisé par une cascade de cellules de second ordre et de fonction de transfert G(Z). Dans le cas autorégressif, il vient :

$$G(z) = a_0 . \prod_{i=1}^{L} \frac{1}{1 + b_1^i Z^{-1} + b_2^i Z^{-2}} \qquad (14.72)$$

Pour contrôler la stabilité d'un tel filtre, il suffit de s'assurer que les conditions suivantes sont remplies, d'après le paragraphe 6.7 :

$$|b_2^i| < 1 ; \quad |b_1^i| < 1 + b_2^i ; \quad 1 \leqslant i \leqslant L ; \qquad (14.73)$$

Comme précédemment, le calcul du gradient de la fonction d'erreur nécessite la connaissance du terme g_k^i tel que :

$$g_k^i(n) = \frac{\partial \tilde{y}(n)}{\partial b_k^i}$$

Considérant que :

$$\tilde{y}(n) = \frac{1}{2\pi j} \int_\Gamma Z^{n-1} . a_0 . \prod_{i=1}^{L} \frac{1}{1 + b_1^i Z^{-1} + b_2^i Z^{-2}} . X(Z)\, dZ$$

on obtient :

$$\frac{\partial \tilde{y}(n)}{\partial b_k^i} = -\frac{1}{2\pi j}\int_\Gamma Z^{n-1}.\frac{Z^{-k}}{1+b_1^i Z^{-1}+b_2^i Z^{-2}}.G(Z).X(Z)\,dZ$$

Cette expression signifie que les termes $g_k^i(n)$, avec $k = 1,2$ et $1 \leqslant i \leqslant L$, sont obtenus en appliquant la suite $\tilde{y}(n)$ à la cellule récursive de rang i. Le schéma correspondant est celui de la figure 14.7, où les cellules de second ordre non récursives sont remplacées par les cellules récursives. Pour chaque valeur de l'indice n la stabilité du système est testée par les relations (14.73).

La méthode qui vient d'être exposée s'applique aussi au modèle ARMA, pour lequel elle conduit à des circuits un peu plus compliqués.

Les techniques utilisées dans les paragraphes précédents procèdent par minimisation globale de l'erreur quadratique moyenne. Il est possible d'obtenir une minimisation séquentielle, étape par étape, en utilisant la structure des treillis.

Le filtre en treillis de la figure (13.3) peut être adapté à chaque valeur de l'indice par un algorithme de gradient ; en effet les signaux en sortie des cellules élémentaires s'écrivent :

$$\begin{aligned} e_i(n) &= e_{i-1}(n) - k_i b_{i-1}(n-1) \\ b_i(n) &= b_{i-1}(n-1) - k_i e_{i-1}(n) \end{aligned} \qquad (14.74)$$

et les gradients ont pour expressions :

$$\begin{aligned} \frac{\partial e_i^2(n)}{\partial k_i} &= -2e_i(n)\,b_{i-1}(n-1) \\ \frac{\partial b_i^2(n)}{\partial k_i} &= -2b_i(n)\,e_{i-1}(n) \end{aligned} \qquad (14.75)$$

Alors, les variations suivantes peuvent être appliquées aux coefficients en supposant que l'on cherche à minimiser les fonctions $E[e_i^2(n) + b_i^2(n)]$, pour $1 \leqslant i \leqslant N$:

$$k_i(n+1) = k_i(n) + \delta_i(e_i(n)\,b_{i-1}(n-1) + b_i(n)\,e_{i-1}(n)) \qquad (14.76)$$

Comme la puissance des signaux $e_i(n)$ et $b_i(n)$ décroît avec l'indice i, le pas de variation δ_i doit être relié à cette puissance pour obtenir l'homogénéité des constantes de temps.

14.9 CONCLUSION

Des techniques pour concevoir et réaliser des filtres adaptatifs ont été présentées. Elles sont basées sur l'algorithme du gradient, qui constitue l'approche la plus

simple et la plus robuste pour faire varier les coefficients. La structure RIF directe a été étudiée en détail, en faisant apparaître les paramètres d'adaptation, constante de temps et erreur résiduelle, et de complexité, cadence des multiplications et nombre de bits des coefficients et des données internes ; c'est la structure la plus utilisée. D'autres structures peuvent, dans certains cas, présenter des avantages intéressants, les structures RII, mixtes RIF-RII ou en treillis. L'analyse des conditions de stabilité de ces structures, l'étude des paramètres d'adaptation et de complexité peuvent se faire en suivant une démarche analogue à celle décrite pour la structure RIF.

L'algorithme du gradient amène une évolution relativement lente des valeurs des coefficients du filtre, notamment quand on recherche une erreur résiduelle faible et quand on l'utilise sous sa forme la plus dépouillée, l'algorithme du signe. Pour atteindre une plus grande rapidité d'adaptation, on peut recalculer périodiquement l'ensemble des coefficients en faisant appel à des procédures itératives rapides ; la structure en treillis se prête bien à cette approche et permet d'analyser ou de modéliser des signaux comme la parole en temps réel avec des circuits de complexité modérée.

L'algorithme du gradient peut être perfectionné, par exemple en utilisant des pas de variation différents et variables pour les coefficients au lieu d'un pas uniforme, obtenus à partir d'estimations des caractéristiques statistiques des signaux.

Des critères mieux appropriés dans certaines applications que la minimisation de l'erreur quadratique moyenne peuvent être envisagés et des algorithmes plus performants que le gradient peuvent être élaborés [1]. Cependant la mise en œuvre de ces algorithmes est généralement nettement plus compliquée et des problèmes de sensibilité aux imperfections de la réalisation peuvent apparaître.

Finalement les structures RIF et RII, fonctionnant suivant le critère de minimisation de l'erreur quadratique moyenne et utilisant l'algorithme du gradient, ou sa forme la plus simple l'algorithme du signe, offrent un bon compromis de simplicité, d'efficacité et de robustesse pour le filtrage adaptatif.

BIBLIOGRAPHIE

[1] M. Bellanger – Adaptative Digital Filters – 2nd edition, Marcel Dekker, Inc., 2001, 464 pages.

[2] O. Macchi – Adaptative Processing, the LMS approach with Application in Transmission – Wiley, New York, 1995.

[3] F. Michaut, M. Bellanger – Filtrage Adaptatif, Théorie et Algorithmes – Lavoisier, 2005.

[4] S. Haykin – Adaptative Filter Theory – 4th édition, Prentice Hall, Englewood Cliffs, New Jersey, 2000.

[5] P. A. Regalia – Adaptative IIR Filterning in Signal Processing and control – Marcel Dekker Inc., 1995.

EXERCICES

1 Le signal $x(n) = m + e(n)$, où m est une constante et $e(n)$ un bruit blanc de puissance σ_e^2 est appliqué à un estimateur récursif, dont la relation entre suite de sortie $y(n)$ et suite d'entrée $x(n)$ est donnée par :

$$y(n) = (1 - b)\,y(n-1) + bx(n)$$

La suite $x(n)$ étant supposée nulle pour $n < 0$, calculer $y(n)$. Si $b = 0{,}8$ combien faut-il d'échantillons pour que $y(n)$ approche m en moyenne à moins de 1 %.

Calculer l'erreur quadratique moyenne en sortie $\mathrm{E}[(y(n) - m)^2]$ pour $n > 0$; quelle est sa limite quand n tend vers l'infini ; étudier son évolution et le choix du coefficient b pour les trois cas suivants : $\sigma_e \simeq m$, $\sigma_e > m$ et $\sigma_e < m$.

Comparer les performances de cet estimateur récursif avec celles de l'estimateur non récursif défini par : $y(n) = \dfrac{1}{n+1} \sum_{i=0}^{n} x(i)$.

2 Un signal sinusoïdal $x(n)$ est appliqué à un filtre prédicteur RIF du second ordre

$$x(n) = \sin 2\pi \frac{3n}{8}$$

Calculer les coefficients a_1 et a_2 de ce filtre et placer les zéros dans le plan des Z.

Les coefficients étant initialement nuls, tracer la trajectoire des zéros de ce filtre pour un pas d'adaptation $\delta = 0{,}1$.

Un bruit blanc discret de puissance σ^2 étant ajouté au signal, calculer les nouvelles valeurs des coefficients de prédiction. Tracer dans ce cas une trajectoire des zéros du filtre.

3 Soit un filtre transversal destiné à égaliser un canal ayant la fonction de transfert :

$$c(z) = \frac{0{,}5}{1 - 0{,}5Z^{-1}}$$

Calculer la puissance du signal reçu $x(n)$ en supposant les données non corrélées et de puissance unité. Pour N = 3 coefficients donner les valeurs de ces coefficients. Un bruit blanc de puissance $\sigma^2_b = 0{,}1$ est ajouté au signal reçu. Calculer les valeurs des coefficients et comparer aux résultats précédents. Quel est le facteur d'amplification du bruit ? À partir de la réponse impulsionnelle de l'ensemble canal-égaliseur, calculer la puissance de l'interférence entre les symboles.

Calculer la valeur propre $\lambda_{\min}$ de la matrice d'autocorrélation R_3 du signal reçu. Dans une réalisation adaptative, on prend $\delta = 0{,}05$ comme pas d'adaptation. Quelle est la constante de temps du filtre adaptatif. Vérifier par simulation.

4 On considère le schéma de la figure 14.3 dans lequel le signal d'entrée est une suite d'échantillons non corrélés et de puissance unité, le filtre $H_{opt}(Z) = \dfrac{1 + 0{,}5Z^{-1}}{1 - 0{,}5Z^{-1}}$,

le filtre adaptatif $H(Z) = \dfrac{a_0 + a_1 Z^{-1}}{1 - b\, Z^{-1}}$ et le bruit blanc additif $b(n)$ a pour puissance $\sigma_b^2 = 0{,}1$.

– Calculer les valeurs optimales des coefficients du filtre adaptatif.
– Écrire les équations de mise à jour dans une réalisation adaptative et donner les bornes du pas d'adaptation δ. Tracer la courbe d'évolution des coefficients pour la valeur $\delta = 0{,}1$, vérifier la constante de temps et l'erreur résiduelle après convergence.

Chapitre **15**

Applications

Le traitement du signal accompagne la généralisation de l'électronique à l'ensemble des secteurs techniques. Quelques exemples d'applications sont présentés dans ce chapitre, principalemnt du domaine des communications.

15.1 DÉTECTION D'UNE FRÉQUENCE

Soit à déterminer l'amplitude de la composante à la fréquence f_0 d'un signal échantillonné à la fréquence $f_e > 2f_0$. La suite des opérations à faire correspond au schéma de la figure 15.1.

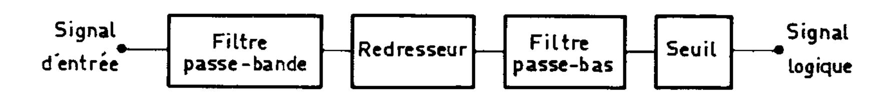


FIG. 15.1. *Détection d'une fréquence par filtrage passe-bande*

Le signal est appliqué à un filtre passe-bande étroit centré sur la fréquence f_0. La fonction de redresseur est ensuite réalisée en prenant la valeur absolue des nombres obtenus. La suite de ces valeurs absolues est appliquée à un filtre passe-bas qui fournit la valeur de l'amplitude cherchée. Si c'est la présence de la composante à la fréquence f_0 qu'il faut détecter, une logique à seuil fournit l'information logique.

Ce traitement peut être analysé comme suit ; soit $s_0(t)$ le signal à détecter avec :

$$s_0(t) = A \sin(\omega_0 t)$$

Prendre la valeur absolue des nombres qui représentent les échantillons de ce signal revient à multiplier ces échantillons par un signal carré $i_p(t)$ en phase et d'amplitude unitaire.

D'après la relation (1.6) du paragraphe 1.1, on peut écrire :

$$i_p(t) = 2 \sum_{n=0}^{\infty} h_{2n+1} \sin [(2n+1)\,\omega_0 t] \qquad (15.1)$$

avec

$$h_{2n+1} = (-1)^n \cdot \frac{\sin\left[\dfrac{\pi(2n+1)}{2}\right]}{\dfrac{\pi(2n+1)}{2}} = \frac{1}{\dfrac{\pi}{2}(2n+1)}$$

Le signal $s_0^*(t)$ obtenu après redressement s'écrit :

$$s_0^*(t) = 2\text{A} \sum_{n=0}^{\infty} h_{2n+1} \sin [(2n+1)\,\omega_0 t] \sin (\omega_0 t)$$

ou encore :

$$s_0^*(t) = \text{A}h_1 + \text{A} \sum_{n=1}^{\infty} (h_{2n+1} - h_{2n-1}) \cos (2n\omega_0 t) \qquad (15.2)$$

Pour obtenir l'amplitude A, il faut éliminer les termes de la somme infinie. Les produits parasites correspondants ont, à partir d'un certain rang, une fréquence supérieure à la demi-fréquence d'échantillonnage $\frac{f_e}{2}$ et se trouvent repliés dans la bande utile. Il faut choisir les caractéristiques du filtre passe-bas, en particulier la fréquence de début de bande affaiblie pour éliminer les parasites les plus importants; ceux qui demeurent en bande passante amènent des fluctuations sur la mesure de l'amplitude A.

Dans cette méthode, il est avantageux d'utiliser un filtre passe bande RII et un filtre passe-bas RIF, car la mesure d'amplitude peut être faite à une cadence inférieure à f_e. Il existe une autre méthode qui permet de n'utiliser que des filtres muticadence; elle est basée sur une modulation par deux porteurs en quadrature à la fréquence f_0 et est présentée à la figure 15.2.

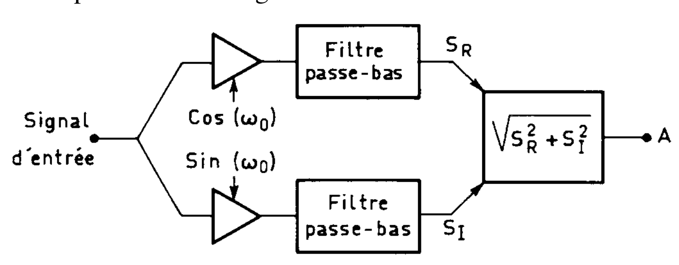


Fig. 15.2. *Détection d'une fréquence par modulation et filtrage passe-bas*

La composante à détecter s'écrit :

$$s(t) = \text{A} \sin (\omega_0 t + \varphi)$$

où φ représente un déphasage de la composante à détecter par rapport au porteur. Après filtrage passe-bas sur les deux branches, pour éliminer les produits parasites, on obtient les signaux :

$$S_R = \frac{A}{2} \sin \varphi; \quad S_I = \frac{A}{2} \cos \varphi \qquad (15.3)$$

L'amplitude cherchée est obtenue par :

$$A = 2\sqrt{S_R^2 + S_I^2}$$

La réalisation rigoureuse du calcul $X = \sqrt{S_R^2 + S_I^2}$ est compliquée et on se contente en général d'une approximation X', qui dépend du déphasage φ.

Le tableau 15.1 donne diverses approximations et les erreurs relatives correspondantes. Ces erreurs peuvent être réduites par multiplication, par un coefficient de recentrage C, c'est-à-dire en calculant la valeur X'_C telle que :

$$X'_C = C\sqrt{S_R^2 + S_I^2}$$

Tableau 15.1. – Approximation de $X = \sqrt{S_R^2 + S_I^2}$

X'	$\max\left(\frac{X' - X}{X}\right)$	C	$\max\left(\frac{X'_C - X}{X}\right)$
$\|S_R\| + \|S_I\|$	0,41421	0,82843	0,17157
Max ($\|S_R\|, \|S_I\|$)	0,29289	1,17157	0,17157
Max ($\|S_R\|, \|S_I\|$) + $\frac{1}{2}$ min ($\|S_R\|, \|S_I\|$)	0,11803	1,05803	0,05573

La détection d'une fréquence avec modulation demande en général moins de calculs que la méthode qui utilise un filtre passe-bande, mais elle nécessite de disposer des signaux porteurs convenables.

La fonction de détection d'une fréquence est utilisée dans les systèmes de transmission des signaux de signalisation et constitue l'essentiel des dispositifs récepteurs de codes multifréquences.

Dans les connexions entre centraux téléphoniques et postes d'abonné, les informations relatives à l'établissement, au maintien et à la tarification des communications sont transmises par un code de signalisation multifréquence. Le poste d'abonné à clavier engendre deux tonalités, l'une provenant d'un groupe de 4 fréquences basses, l'autre d'un groupe de 4 fréquences hautes, constituant ainsi un ensemble de 16 codes possibles. Les valeurs de ces fréquences sont données dans le tableau 15.2.

Tableau 15.2. – Fréquences du code de clavier

Code de clavier	Fréquences en Hz			
BF	697	770	852	941
HF	1209	1336	1477	1633

Le récepteur de code multifréquence a pour fonction de détecter la présence de ces deux tonalités et de fournir l'indication du code représenté.

15.2 BOUCLE À VERROUILLAGE DE PHASE

Cette fonction intervient pour la récupération de rythme dans les terminaux et les récepteurs [1-2]. Le principe est illustré à la figure 15.3.

Quand la boucle est à l'équilibre, la fréquence produite par l'oscillateur contrôlé en tension est égale à la fréquence du signal d'entrée et le détecteur de phase produit un signal dont la composante continue est extraite par le filtre passe-bas étroit. Le détecteur de phase peut être un modulateur qui effectue le produit de la sortie de l'oscillateur par le signal d'entrée. Si la fréquence nominale de l'oscillateur est égale à la fréquence d'entrée, les signaux sont en quadrature et la composante continue en sortie du modulateur est nulle. Si ce n'est pas le cas, l'écart de phase par rapport à la quadrature produit une composante continue qui décale la fréquence de l'oscillateur de la quantité nécessaire pour que les fréquences soient égales. La largeur de bande du filtre de boucle détermine la plage d'accrochage de l'asservissement, le temps de réaction et le niveau de bruit résiduel.

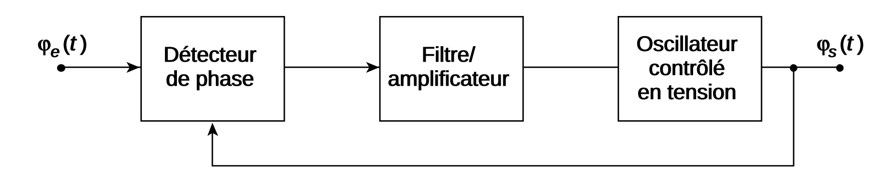


Fig. 15.3. *Principe d'une boucle à verrouillage de phase*

En numérique, on peut reproduire intégralement ce fonctionnement. Cependant, il existe une souplesse supplémentaire, dans la mesure où on peut effectuer le calcul des phases. L'oscillateur numérique peut être réalisé par un accumulateur de phase connecté à une mémoire qui fournit les échantillons de la sinusoïde. On peut donc traiter directement les valeurs des phases à l'entrée et à la sortie de la boucle et l'écart de phase peut être obtenu par simple soustraction. Un modèle correspondant à une boucle du second ordre est donné à la figure 15.4. C'est un asservissement à 2 coefficients K_1 et K_2, correspondant respectivement à la commande dite proportionnelle et intégrale.

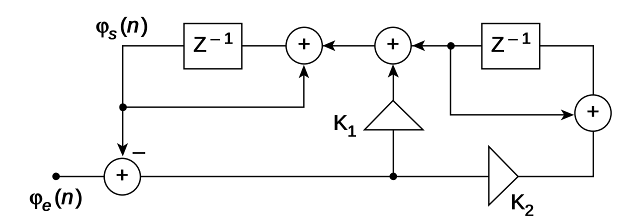


FIG. 15.4. *Modèle d'une boucle du second ordre*

L'oscillateur contrôlé en tension est représenté par l'intégrateur qui fournit la phase de sortie $\varphi_s(n)$. La fonction de transfert entre la sortie et l'entrée s'écrit :

$$H(Z) = \frac{\Phi_s(Z)}{\Phi_e(Z)} = \frac{K_1 Z^{-1} + (K_2 - K_1) Z^{-2}}{1 - (2 - K_1) Z^{-1} + (1 - K_1 + K_2) Z^{-2}} \qquad (15.4)$$

C'est une fonction de transfert de filtre passe-bas, qui vaut un à la fréquence zéro. Les caractéristiques du filtre sont déterminées par les deux paramètres K_1 et K_2. Le domaine de stabilité s'étudie en utilisant les résultats du paragraphe VI.7 et en posant $b_1 = K_1 - 2$ et $b_2 = 1 - K_1 + K_2$. Il vient alors : $1 - K_1 + K_2 < 1$; $|K_1 - 2| < 2 - K_1 + K_2$.

Dans le plan des coefficients K_1 et K_2, le domaine de stabilité est un triangle.

La fonction de transfert entre l'écart de phase et l'entrée s'écrit :

$$\frac{\Phi_e(Z) - \Phi_s(Z)}{\Phi_e(Z)} = \frac{(1 - Z^{-1})^2}{1 - (2 - K_1) Z^{-1} + (1 - K_1 + K_2) Z^{-2}} \qquad (15.5)$$

La présence du terme $(1 - Z^{-1})^2$ au numérateur indique qu'une telle boucle est capable de suivre une dérive de phase.

15.3 CODAGE MIC-DIFFÉRENTIEL

La compression des signaux est une des applications de la prédiction linéaire. Une illustration est donnée dans le domaine de la téléphonie par le codage Mic-différentiel [3].

Pour une voie téléphonique le codage à 8 bits, avec une fréquence d'échantillonnage de 8 kHz, conduit au débit de 64 kbit/s. Le signal peut être à spectre plat ou sinusoïdal avec l'amplitude maximale admissible dans toute la gamme des fréquences utiles ; la distorsion de quantification a une distribution spectrale considérée comme uniforme. Or le signal de parole présente une densité spectrale qui, à long terme, décroît fortement avec la fréquence à partir d'une valeur inférieure à 1 kHz. Dans ces conditions une adaptation du codage aux caractéristiques du signal permet une réduction du débit binaire.

Le principe du codage MIC-différentiel repose sur une prédiction linéaire du signal à coder, comme indiqué au paragraphe 13.5. C'est la différence $e(n)$ entre le

signal $x(n)$ et la prédiction $\tilde{x}(n)$ qui est codée et transmise à chaque période d'échantillonnage. Comme le signal d'erreur $e(n)$ a une distribution spectrale proche d'une distribution uniforme, il exploite beaucoup mieux les possibilités du codeur que le signal de parole initial.

La figure 15.5 donne le schéma de principe du codage MIC-différentiel. La suite $e'(n)$ résulte de l'addition à $e(n)$ de l'erreur de quantification et $x'(n)$ est la suite issue du décodeur. Le signal $e(n)$ a pour expression :

$$e(n) = x(n) - \tilde{x}(n) = x(n) - \sum_{i=1}^{N} a_i x(n-i)$$

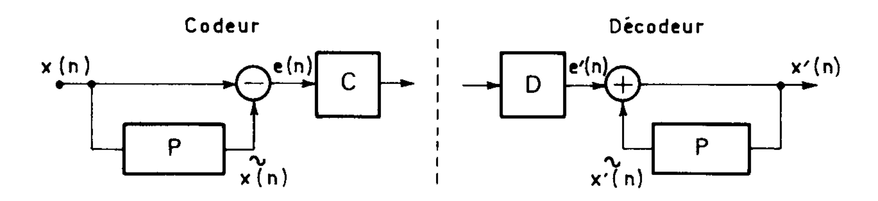


FIG. 15.5. *Principe du codage MIC-différentiel*

Le filtre de prédiction P a pour fonction de transfert la fonction $P(z)$ telle que :

$$P(Z) = \sum_{i=1}^{N} a_i Z^{-1} \tag{15.6}$$

L'ordre du filtre N et les coefficients $a_i (1 \leqslant i \leqslant N)$ doivent être choisis pour minimiser la puissance du signal $e(n)$. Dans ces conditions, pour une valeur donnée de N, les coefficients sont calculés par l'expression 13.42, qui fait intervenir les éléments $r(k)(0 \leqslant k \leqslant N)$ de la fonction d'autocorrélation du signal $x(n)$. Pour le signal de parole les évaluations suivantes ont été proposées :

$$r(0) = 1\,; \quad r(1) = 0{,}8644\,; \quad r(2) = 0{,}5570\,; \quad r(3) = 0{,}2274$$

Elles mettent en évidence une forte corrélation entre échantillons voisins. Les coefficients correspondants ont pour valeur :

$$a_1 = 1{,}936\,; \quad a_2 = -1{,}553\,; \quad a_3 = 0{,}4972$$

Les valeurs propres de la matrice d'autocorrélation R_3 sont :

$$\lambda_1 = 2{,}532\,; \quad \lambda_2 = 0{,}443\,; \quad \lambda_3 = 0{,}025$$

Comme :

$$A_{opt}^t R_3 A_{opt} = 0{,}947$$

le gain de prédiction correspondant est voisin de 13 dB.

Dans la réalisation des systèmes MIC-différentiel pour les télécommunications, un certain nombre d'améliorations ont été apportées au principe de la figure 15.6, pour atteindre un degré de qualité élevé avec des équipements peu complexes :

• la prédiction est faite en utilisant la suite $e(n)$ transmise après quantification, ce qui réduit la puissance de distorsion de quantification. De plus l'émetteur et le récepteur fonctionnent à partir des mêmes informations et il n'est pas nécessaire de transmettre des informations supplémentaires, par exemple pour introduire des procédures adaptatives.

• Le quantificateur est rendu adaptatif en faisant varier l'échelon de quantification en fonction d'une estimation de la puissance du signal. Cette approche permet de tirer profit du fait que les signaux téléphoniques ont une puissance soit constante soit lentement variable. La parole par exemple est un signal non stationnaire qui, cependant, peut être considéré comme stationnaire sur des durées courtes, de l'ordre de 10 ms.

• Le prédicteur est rendu adaptatif pour tenir compte des spectres des divers signaux et suivre les évolutions à court terme du spectre de la parole.

Ces systèmes sont de type MIC-Différentiel Adaptatif (MIC-DA).

Le schéma du filtre de prédiction dans le codeur et du filtre de reconstruction dans le décodeur est donné par la figure 15.6. La fonction de transfert $\mathrm{H}(\mathrm{Z})$ du codeur est donnée par :

$$\mathrm{H}(\mathrm{Z}) = 1 - \sum_{i=1}^{\mathrm{N}} a_i \mathrm{Z}^{-i} = \frac{1}{1 + \dfrac{\sum_{i=1}^{\mathrm{N}} a_i \mathrm{Z}^{-i}}{1 - \sum_{i=1}^{\mathrm{N}} a_i \mathrm{Z}^{-i}}} \tag{15.7}$$

La prédiction $\tilde{x}(n)$ est calculée à partir de la suite reconstituée $y(n)$. Le quantificateur Q introduit une distorsion qui s'ajoute au signal d'entrée et qui est supposée avoir un spectre uniforme. Par rapport au principe de la figure 15.5, la distorsion de quantification est réduite d'un facteur approchant le gain de prédiction G.

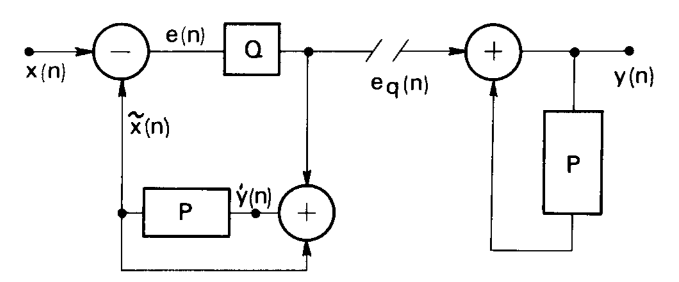


FIG. 15.6. *Filtres dans un système MIC-différentiel*

Le rapport signal à bruit SB du système complet s'écrit :

$$SB = \frac{\sum_n x^2(n)}{\sum_n [y(n) - x(n)]^2} = \frac{\sum_n x^2(n)}{\sum_n e^2(n)} \cdot \frac{\sum_n e^2(n)}{\sum_n [y(n) - x(n)]^2}$$

Soit pour des signaux stationnaires :

$$SB = \frac{E[x^2(n)]}{E[e^2(n)]} \cdot \frac{E[e^2(n)]}{E[(y(n) - x(n))^2]} \qquad (15.8)$$

Le premier terme est le gain de prédiction. Si le signal d'erreur $e(n)$ a un spectre uniforme et si la distorsion de quantification $q(n) = e(n) - e_q(n)$ est faible, alors le gain de prédiction G est tel que :

$$G^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{d\omega}{\left| 1 - \sum_{i=1}^{N} a_i e^{-ji\omega} \right|^2} \qquad (15.9)$$

Les égalités suivantes sont vérifiées :

$$q(n) = e(n) - e_q(n) = x(n) - \tilde{x}(n) - e_q(n) = x(n) - y(n)$$

Dans le cas d'une transmission sans erreurs entre codeur et décodeur, le rapport signal à bruit du système s'exprime en décibels par :

$$SB = 20 \log G + 10 \log \frac{E[e^2(n)]}{E[q^2(n)]} \qquad (15.10)$$

Pour maximiser le deuxième terme de cette expression, un quantificateur optimal peut être utilisé, comme au paragraphe 1.12.

La fonction de transfert H(Z) du codeur peut être de type RIF, RII ou mixte RIF-RII. En utilisant les techniques du chapitre 14 les filtres de prédiction et reconstitution peuvent être rendus adaptatifs.

Le schéma d'un transcodeur entre une voie MIC au débit de 64 kbit/s et une voie MIC-DA à 32 kbit/s est donné à la figure 15.7. L'échelon de quantification est ajusté en fonction de la racine carrée de la puissance moyenne du signal $x(n)$, estimée à partir de la sortie du quantificateur.

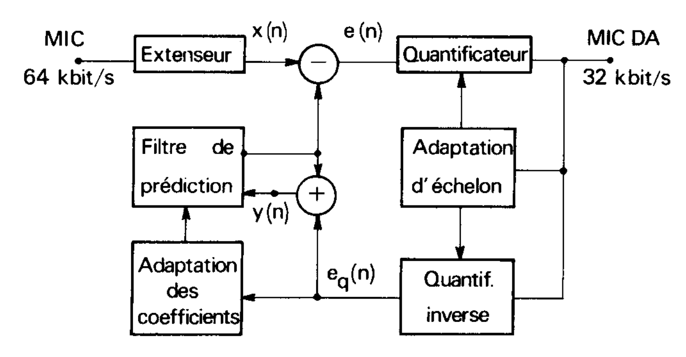


FIG. 15.7. *Schéma de transcodeur MIC-MICDA*

La qualité de transmission de ce système est telle qu'il ne produit aucune dégradation notable pour la plupart des signaux téléphoniques. Avec un prédicteur d'un ordre suffisant, par exemple RIF d'ordre 10, le gain de prédiction sur la parole va de 6 dB pour les sons non voisés à 16 dB pour les sons voisés, avec une valeur subjective globale de 13 dB environ; pour des signaux de modems rapides, 4800 bit/s, il est d'environ 4 dB. Le rapport signal à bruit du quantificateur optimal à 4 bits est voisin de 20 dB. Avec ces valeurs, le rapport signal à bruit d'un système MIC-DA est suffisant pour transmettre la parole est les données dans de bonnes conditions.

Sur le plan international un algorithme de conversion MIC-MICDA a été normalisé par l'UIT (Union Internationale de Télecommunications); il est décrit dans l'avis G 721 [4].

15.4 CODAGE DU SON

Les bancs de filtres sont à la base de la compression numérique du son, car ils permettent de profiter des particularités de l'oreille et notamment de l'effet de masquage.

L'algorithme représenté à la figure 15.8 et normalisé sous l'appellation UIT-T/G722, permet la transmission du son sur le canal téléphonique à 64 kbit/s. Le signal audio a une bande dite élargie de 7 kHz, il est échantillonné à 16 kHz et codé à 14 bits. Un banc de 2 filtres QMF permet d'obtenir 2 sous-bandes échantillonnées à 8 kHz qui sont ensuite codées en MIC-DA aux débits de 48 et 16 kbit/s pour les bandes basse et haute respectivement. Une opération de multiplexage, avec insertion de données éventuellement, fournit le débit de 64 kbit/s à transmettre.

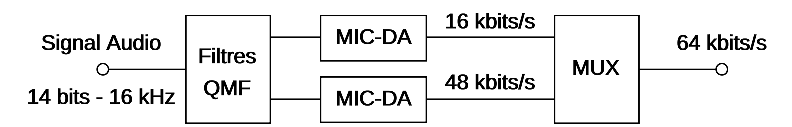


FIG. 15.8. *Codage à deux sous-bandes d'un signal audio*

La compression du son de haute qualité pour la diffusion numérique ou le stockage est définie par la norme ISO/CEI 11172 [5-6]. Elle est basée sur un banc de 32 filtres à 512 coefficients, de type pseudo-QMF. Les signaux ainsi obtenus sont quantifiés séparément, avec un nombre de bits pour chaque sous-bande tel que le bruit de quantification reste à un niveau inférieur au seuil de masquage. Ce seuil, illustré à la figure 15.9, est défini pour chaque sous-bande à partir d'une analyse par TFD à 1 024 points en appliquant les résultats de la psycho-acoustique.

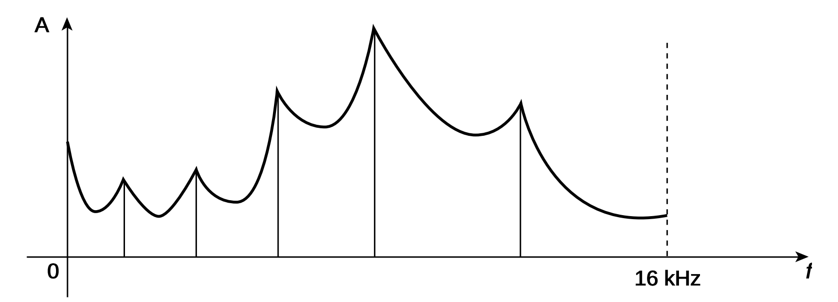


FIG. 15.9. *Exemple de courbe de masquage pour un signal sonore*

La méthode permet d'atteindre un débit de 64 kbit/s pour une voie son monophonique de haute qualité. (MPEG – Couche 3 / MP3) [6].

15.5 ANNULATION D'ÉCHO

Dans les réseaux de transmission, des échos se produisent quand une réplique retardée et affaiblie du signal, émis par un terminal local à destination d'un terminal distant, atteint le récepteur local. De tels signaux d'échos ont leur origine dans les transformateurs hybrides qui effectuent la conversion 2 fils - 4 fils, dans les désadaptations d'impédance le long des lignes à 2 fils et, dans certains cas particuliers comme la téléphonie mains-libres, dans les couplages acoustiques entre les haut-parleurs et les microphones des terminaux. L'annulation d'écho consiste à modéliser ces couplages parasites entre émetteurs et récepteurs locaux et à soustraire un écho synthétique de l'écho réel [7].
Deux situations différentes peuvent être distinguées, selon la nature des signaux impliqués, à savoir la parole et les données. Le cas des modems pour transmission de données est traité en premier, car il est plus simple à aborder.

15.5.1 Annuleur d'écho pour données

L'exploitation la plus efficace des lignes à 2 fils est réalisée quand les signaux de données sont transmis simultanément dans les deux directions et dans les mêmes bandes de fréquence. La transmission est alors dite bi-directionnelle ou full-duplex. Le principe est illustré à la figure 15.10.
Le signal $x_A(n)$ est transmis du terminal A au terminal B à travers une ligne à 2 fils. Le signal $y(n)$ à l'entrée du récepteur du terminal A possède 2 composantes, le signal y_B(n) provenant du terminal B qui est le signal de données utile et le signal réfléchi qui constitue l'écho perturbateur produit par $x_A(n)$ et désigné par $r_A(n)$. La fonction du filtre $H(z)$ est de produire un écho synthétique $y(n)$ aussi proche que possible de $r_A(n)$, de sorte que, après soustraction, le signal d'erreur $e(n)$ reste suffisamment proche de $y_B(n)$ pour que la transmission des données du terminal A au terminal B ait la qualité suffisante.
Le choix des paramètres du filtre adaptatif est guidé par le contexte. Le nombre N des coefficients est déduit de la durée de la réponse impulsionnelle d'écho à compenser, en tenant compte de la fréquence d'échantillonnage.

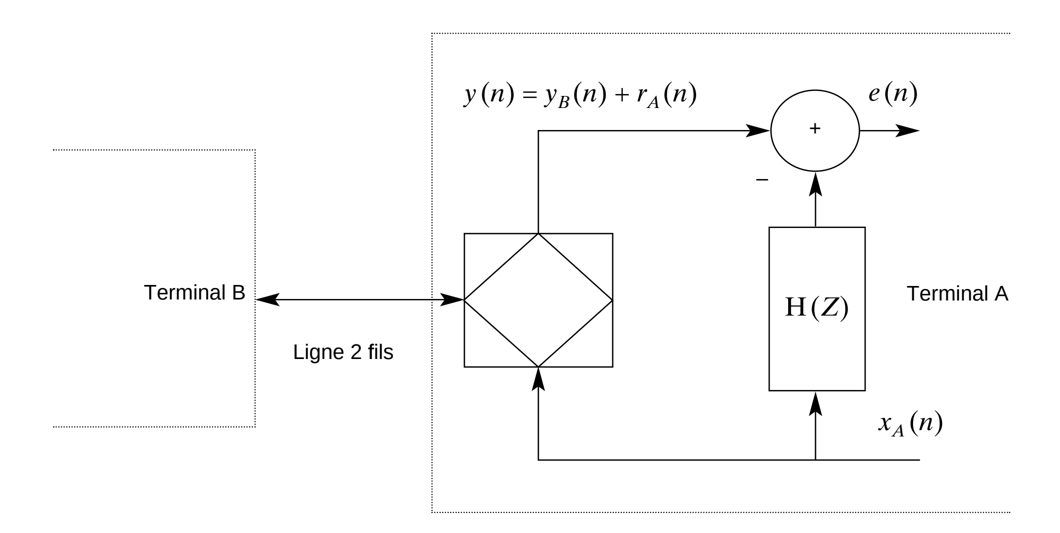


FIG. 15.10. Principe de l'annulation d'écho

Il est nécessaire de rendre le filtre adaptatif, parce que les caractéristiques de la ligne de transmission peuvent changer avec le temps. Pour tout filtre adaptatif, la nature des signaux d'entrée est importante et dans le cas présent la situation est très favorable. En effet, comme le montre la figure 15.10, l'entrée du filtre est le signal de données émis $x_A(n)$, qui est généralement non corrélé, a une puissance unité et donc possède la matrice d'autocorrélation $R_N = I_N$. Alors, l'algorithme du gradient a des performances équivalentes à celle de l'algorithme des moindres

carrés, le pas d'adaptation δ étant borné par 2/N et la constante de temps a pour valeur $\tau = 1/\delta$. Dans la phase d'apprentissage, la puissance moyenne de l'erreur de sortie s'écrit, en appliquant la relation (14.28) :

$$E_r(n) = \|Hopt\|_2^2(1 - \delta)^{2n} \quad (15.11)$$

La norme L_2 du vecteur des coefficients d'écho $\|H_{opt}\|_2^2$, représente la puissance du signal d'écho.
En transmission bidirectionnelle, le signal utile $y_B(n)$ dans la référence est plus faible que l'écho $r_A(n)$. Si A_s désigne le rapport de l'écho au signal utile, SB le rapport du signal désiré au bruit à l'entrée du récepteur, alors l'affaiblissement d'écho A_e doit satisfaire l'inégalité suivante :

$$A_e > A_s + SB\ (dB) \quad (15.12)$$

Par exemple, avec SB = 40 dB et A_s = 20 dB, alors A_e = 60 dB, ce qui implique un haut niveau de performance de l'annuleur d'écho, pour ce qui est de l'erreur résiduelle en régime permanent, après convergence.
Dans l'adaptation, le signal utile crée un écart sur les coefficients par rapport à la valeur optimale et il en résulte une augmentation de l'erreur résiduelle en sortie. En désignant par σ_y^2 la puissance du signal utile des données reçues et en utilisant les résultats de l'étude de l'algorithme du gradient, la variance de chaque coefficient de filtre après convergence a pour valeur $\sigma_y^2\delta/2$ et l'erreur résiduelle est N fois plus grande, soit N $\sigma_y^2\delta/2$. Pour atteindre un objectif de rapport signal à bruit SB, le pas d'adaptation doit satisfaire l'inégalité :

$$N\frac{\delta}{2} < \frac{1}{SB} \quad (15.13)$$

Dans cette approche, on a supposé que la puissance de l'erreur de sortie est voisine de celle du signal utile σ_y^2. Par exemple, avec SB = 10^4 (40 dB) et N = 60, il faut prendre $\delta < 3{,}3\ 10^{-6}$. C'est une valeur très faible, qui entraine une très longue phase d'apprentissage. Il est important de noter l'impact sur la précision des coefficients. En appliquant les résultats sur la précision des coefficients en filtrage adaptatif par gradient, on obtient l'expression simplifiée suivante pour le nombre de bits b_c des coefficients :

$$b_c = \log 2(\frac{1}{\delta}) + \frac{1}{2}\log_2(A_e) \quad (15.14)$$

Avec $\delta = 3{,}3\ 10^{-6}$ et $A_e = 10^6$, il vient : $b_c \approx 29$. En pratique, il n'est pas obligatoire d'utiliser cette précision dans les multiplications du filtre, elle est seulement nécessaire dans la mise à jour des coefficients.

15.5.2 L'annuleur d'écho acoustique

L'annulation d'écho acoustique conduit à des filtres de très grande longueur. Avec la vitesse de propagation des ondes acoustiques dans l'air v = 330 m/s, une fréquence d'échantillonnage f_e = 8 kHz et une distance parcourue 2D = 100 m on aboutit à $N = \frac{2D}{v} f_e = 2400$.

On rencontre cette situation dans les salles d'audioconférence par exemple.

De plus, l'annuleur d'écho doit faire face à des situations difficiles pour l'adaptativité, comme la double parole. Durant la conversation, il peut arriver que les 2 utilisateurs parlent en même temps et il y a transmission bidirectionnelle simultanée. Cette situation amène des écarts sur les coefficients, ce qui réduit l'affaiblissement d'écho. En fait, pendant la double parole, il faut geler les coefficients, ce qui pose le problème de la détection de double parole. Une approche simple est indiquée à la figure 15.11. Elle consiste à comparer le niveau du signal reçu r (n) avec

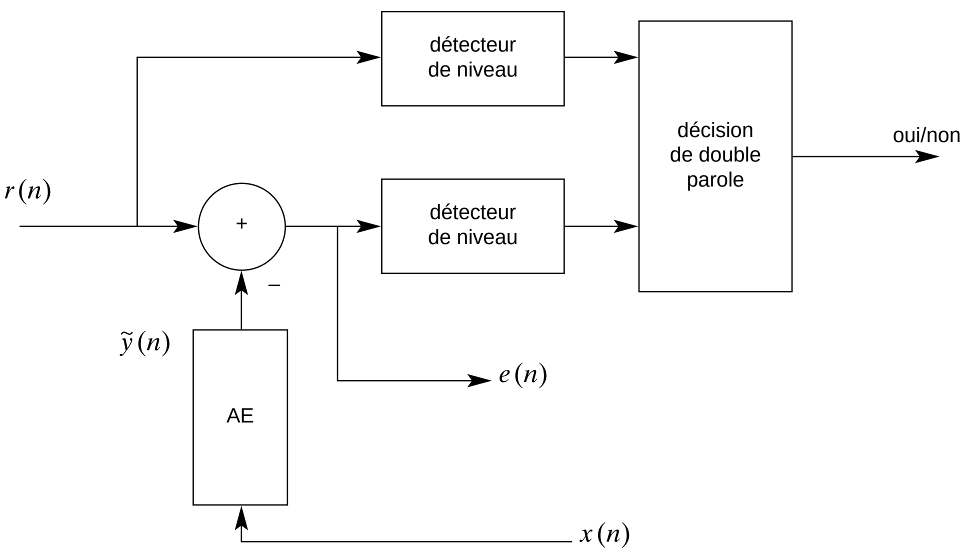


Fig. 15.11. *Détection de double parole*

le niveau du signal e(n) après soustraction de l'écho synthétique. En l'absence de signal de parole distante dans r(n) et si l'annulation fonctionne correctement, les niveaux sont nettement différents. Au contraire, pendant la double parole, les niveaux se rapprochent. Avec cette information, on peut décider de la présence de double parole et geler les coefficients.

Les paramètres pour la détection de niveau et la décision doivent être choisis avec soin, pour éviter une fausse décision et des retards excessifs. Les détecteurs de niveau peuvent être basés sur des mesures de puissance ou d'amplitude.

15.6 TRAITEMENT DES IMAGES DE TÉLÉVISION

Les vidéocommunications et la diffusion des signaux audiovisuels en numérique s'appuient sur des techniques de compression d'images à base de traitement à une ou deux dimensions.

Une image animée est une fonction $s(x, y, t, \lambda)$, de 4 variables, qui sont les deux variables du plan, le temps et la longueur d'onde. En télévision, ce signal est ramené à une seule dimension pour la transmission.

La variable longueur d'onde peut être retirée en considérant que le système visuel humain comporte trois types de récepteurs qui effectuent des fonctions de filtrage conduisant à 3 signaux correspondant aux couleurs primaires Rouge, Vert et Bleu (R, V, B).

Le balayage de télévision ramène ensuite ces signaux à trois dimensions à des signaux monodimensionnels. Les images sont analysées 25 fois par seconde, à raison de 625 lignes par analyse. En fait les lignes paires et impaires d'une image sont regroupées dans deux trames différentes multiplexées dans le temps, d'où une fréquence de 50 trames entrelacées par seconde.

Pour la transmission, les composantes primaires R, V, B sont remplacées par des combinaisons linéaires, appelées respectivement luminance Y et différences de couleur ou chrominance Dr et Db, telles que :

$$\begin{aligned} \mathrm{Y} &= \quad 0{,}30\mathrm{R} + 0{,}59\mathrm{V} + 0{,}11\mathrm{B} \\ \mathrm{Dr} = \mathrm{R} - \mathrm{Y} &= \quad 0{,}70\mathrm{R} - 0{,}59\mathrm{V} + 0{,}11\mathrm{B} \\ \mathrm{Db} = \mathrm{B} - \mathrm{Y} &= -0{,}30\mathrm{R} - 0{,}59\mathrm{V} + 0{,}89\mathrm{B}. \end{aligned}$$

La numérisation de ces signaux se fait avec une fréquence de 13,5 MHz pour la luminance et de 6,75 MHz pour les signaux de chrominance. La conversion Analogique/Numérique étant à 8 bits, le débit correspondant s'élève à 216 Mbit/s. Ce format correspond à la recommandation CCIR 601 de l'UIT et il est dit de type 422. Il conduit à des images se présentant sous la forme de tableaux de nombres de 8 bits comprenant 720 points par ligne et 576 lignes utiles, dans le cas d'un balayage à 625 lignes.

Une image correspond donc à 414 720 octets pour la luminance et 207 360 octets pour chacune des composantes de chrominance.

Les techniques de réduction de débit s'appuient sur le fait qu'une bonne modélisation est fournie par la sortie d'un filtre RII du premier ordre auquel est appliqué un bruit blanc gaussien [8]. La fonction d'autocorrélation à 2 dimensions correspondante s'écrit :

$$\mathrm{V}(x, y) = r_0 e^{-(\alpha x + \beta y)}$$

où α et β sont des constantes positives. Pour le spectre associé, il vient :

$$S(\omega_1, \omega_2) = r_0 \frac{4\alpha\beta}{(\alpha^2 + \omega_1^2)(\beta^2 + \omega_2^2)} \tag{15.15}$$

La plus grande compression dans la représentation d'un signal est obtenue par la transformation propre, basée sur les vecteurs propres de la matrice d'autocorrélation. Dans le cas des signaux du premier ordre, cette transformation est bien approchée par les transformées en cosinus ou sinus discrètes, présentées aux paragraphes 3.3.3 et 3.3.4.

Dans les normes de compression d'images, c'est la TCD, appliquée à des blocs de 8×8 points d'image élémentaires ou « pixels », qui a été retenue.

Les normes élaborées pour la visiophonie, le stockage des images et la télévision numérique font appel à la combinaison des 3 techniques suivantes [6] :

– l'estimation du mouvement, pour pouvoir minimiser la différence entre l'image courante et l'image précédente ;

– la transformation en cosinus discrète pour minimiser la redondance spatiale ;

– le codage statistique à longueur variable (CLV).

Le schéma général d'un codeur d'image est donné à la figure 15.12. Le quantificateur Q opère à partir de seuils qui peuvent être pilotés par un dispositif de régulation, permettant à l'aide d'une mémoire tampon d'atteindre un débit constant.

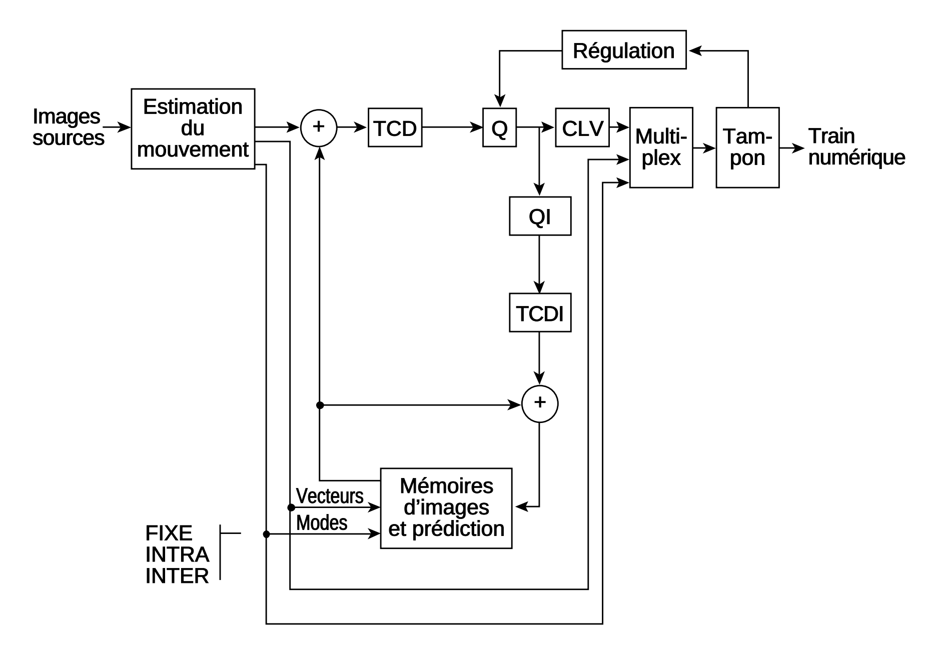


FIG. 15.12. *Schéma général d'un codeur d'images animées*

Pour la télévision de qualité commerciale le débit peut être réduit jusqu'à moins de 4 Mbit/s soit un facteur de compression de l'ordre de 50 [9].

Des filtres numériques interviennent dans les opérations d'interpolation ou de sous-échantillonnage, par exemple pour les changements de formats d'image ou l'estimation de mouvement. Ce sont des filtres séparables.

La compression numérique des signaux multimédia, parole, image et son, permet de réduire considérablement les débits nécessaires à la diffusion des programmes et elle offre la possibilité, associée aux techniques de transmission numérique à haute efficacité spectrale, d'émettre plusieurs programmes dans des canaux utilisés en analogique pour un seul programme.

Les techniques à haute efficacité spectrale font une utilisation intensive du traitement numérique et elles tirent le meilleur profit des particularités des canaux. C'est ainsi que les techniques multiporteuses peuvent conduire à des débits de plusieurs bits/s/Hz sur des canaux de qualité limitée ou susceptibles d'être perturbés.

15.7 TRANSMISSION MULTIPORTEUSE – OFDM

L'objectif des techniques de transmission multiporteuses est d'approcher la capacité théorique d'un canal, d'une part en limitant l'effet des distorsions, d'autre part en ajustant le débit à la densité spectrale de bruit. En effet, en divisant un canal donné en plusieurs dizaines, centaines ou milliers de sous-canaux, on rend négligeable l'effet des distorsions sur chaque sous-canal et on peut affecter à chacun le débit qu'il est capable de supporter. Une approche simple et efficace pour mettre en œuvre ce principe consiste à faire appel à la Transformation de Fourier Rapide, c'est la technique dite OFDM (Orthogonal Frequency Division Multiplexing) dont le principe est représenté à la figure 15.13. Le flux de données à transmettre est converti en N flux élémentaires à débit N fois plus faible, qui sont appliqués à l'entrée d'un calculateur de TFD inverse. Conformément à la définition de la TFD inverse donnée au chapitre 2, cette opération correspond à une modulation par les flux élémentaires de N porteuses, aux fréquences multiples de $\frac{f_e}{N}$ et à l'addition de l'ensemble des signaux ainsi modulés. La cadence des symboles OFDM est alors de $\frac{f_e}{N}$. A la réception, après passage par le canal, une TFD directe effectue l'ensemble des démodulations et restitue les données d'origine, qu'il suffit de sérialiser pour retrouver le flux total initial.

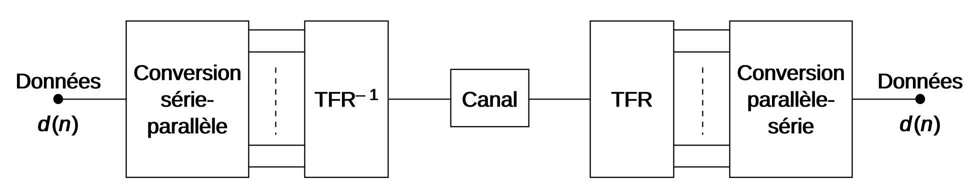


FIG. 15.14. *Principe de la transmission OFDM*

Ce principe très simple est l'illustration directe de la définition de la TFD et de son inverse. Cependant, pour fonctionner correctement, il nécessite plusieurs précautions et adaptations.[10]

En se reportant au paragraphe 2.4 et, en particulier, à la figure 2.8, on observe que l'orthogonalité des signaux n'est vérifiée que pour les fréquences qui sont au centre de l'intervalle de largeur $\frac{f_e}{N}$ affecté à chaque sous-canal, que les sous-canaux ont un domaine de recouvrement et que l'amplitude du recouvrement se réduit avec l'éloignement en fréquence. Sur les bords du canal de transmission, les réponses en fréquence des sous-canaux sont non symétriques, ce qui entraîne des interférences. Il faut donc éviter d'utiliser les sous-canaux extrêmes et prévoir une marge d'au moins quelques sous-canaux, de chaque côté de la bande de fréquence utilisée.

Dans le domaine temporel, un canal de transmission réel a une réponse impulsionnelle de durée τ. Pour qu'il n'y ait aucune superposition entre deux symboles OFDM consécutifs à la réception, il faut que les symboles soient séparés d'une durée suffisante, c'est-à-dire qu'il faut introduire un temps de garde Tg, tel que $Tg > \tau$. Pendant la durée de ce temps de garde, il faut prolonger le symbole OFDM, pour introduire la convolution circulaire indiquée au paragraphe 2.1 et éviter ainsi des interférences entre les sous-canaux. Dans la pratique, c'est la fin du symbole, sur une durée Tg qui est reproduite au début, comme indiqué sur la figure 15.14, ce qui facilite le fonctionnement du récepteur.

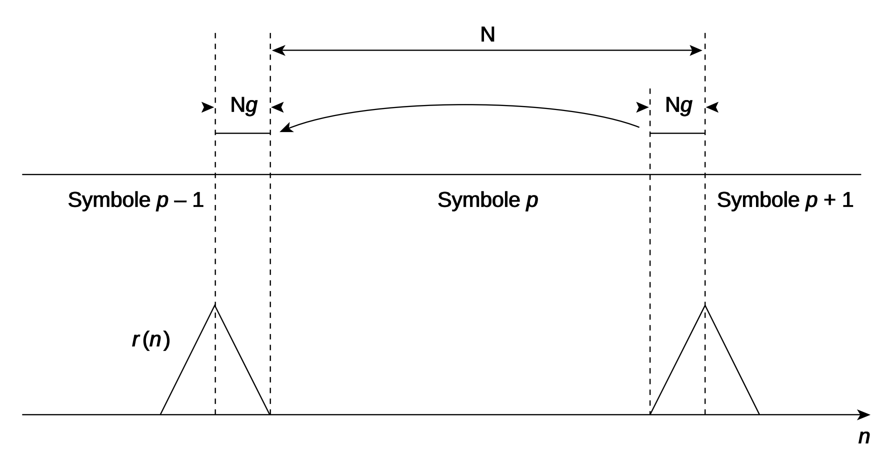


FIG. 15.14. *Introduction du temps de garde et fonction de corrélation*

Avec ce dispositif, les signaux reçus se trouvent simplement multipliés par la TFD du canal, effet qui peut être compensé par une égalisation en amplitude et en phase dans chaque sous-canal. Pour le montrer, on définit par $C(Z)$ la fonction de transfert en Z du canal qui comprend $P + 1$ coefficients, avec $P \leqslant N_g$:

$$C(Z) = \sum_{p=0}^{P} C_p Z^{-p} \tag{15.16}$$

Si $x(n)$ est le signal émis, le signal reçu $y(n)$ s'écrit :

$$y(n) = \sum_{p=0}^{P} C_p x(n-p)$$

Comme $x(n)$ s'exprime à partir des données d_k par

$$x(n) = \sum_{k=0}^{N-1} d_k e^{+j\frac{2\pi}{N}kn} \quad (15.17)$$

On obtient pour $y(n)$ la double sommation

$$y(n) = \sum_{p=0}^{P} C_p \sum_{k=0}^{N-1} d_k e^{j\frac{2\pi}{N}k(n-p)} \quad (15.18)$$

En posant :

$$H_k = \sum_{p=0}^{P} C_p e^{-j\frac{2\pi}{N}kp}$$

il vient finalement

$$y(n) = \sum_{k=0}^{N-1} (d_k H_k) e^{j\frac{2\pi}{N}kn} \quad (15.19)$$

On retrouve bien la propriété de convolution circulaire de la TFD et le récepteur fournit les données émises multipliées par le spectre du canal H_k.

La redondance des signaux émis peut être exploitée par le récepteur pour la synchronisation. En effet, en calculant la fonction de corrélation suivante :

$$r(n) = \sum_{i=n-N_g+1}^{n} y(i)\, y^*(i-N) \quad (15.20)$$

on fait apparaître des pics, comme indiqué sur la figure 15.14, qui caractérisent le début de chaque symbole, permettent de caler la fenêtre d'analyse temporelle du récepteur et peuvent contribuer à la synchronisation des horloges.

La synchronisation en temps et en fréquence est un problème délicat dans les systèmes à grand nombre de porteuses et des symboles particuliers de référence sont introduits ou certains sous-canaux sont réservés à des signaux fixes appelés pilotes.

La figure 15.15 représente le schéma par bloc d'un récepteur de télévision numérique pour la diffusion terrestre [11]. Les interfaces analogiques amènent le signal dans la bande 0,76 – 8,37 MHz et la conversion Analogique-Numérique s'effectue à f_e = 18,28 MHz. Ensuite, une conversion réel-complexe par filtre de quadrature est effectuée et la fréquence d'échantillonnage est ramenée à 9,14 MHz. Avant le calcul de la TFR à N = 8192 points, un multiplieur complexe effectue le calage en fréquence du spectre du signal. La synchronisation temporelle commande le positionnement de la fenêtre de la TFR. Le signal émis comporte 6817 porteurs actifs dont

177 sont consacrés à des signaux pilotes, ce qui permet une synchronisation fine du récepteur, une estimation de la réponse fréquentielle du canal pour l'égalisation et une mesure de la distorsion dans chaque sous-canal. Le temps de garde peut atteindre jusqu'à 20 % de la durée du symbole.

Ce système doit permettre la transmission de débits jusqu'à 32 Mbit/s dans un canal avec espacement de 8 MHz, soit 4 bits/s/Hz.

Un autre exemple d'application de l'OFDM est le système ADSL (Asymmetric Digital Subscriber Line) qui permet la transmission vers l'abonné sur sa paire de cuivre de débits pouvant atteindre 6 Mbit/s au-dessus du signal téléphonique. Le signal multiporteuse occupe une bande de 1 MHz avec 256 porteuses.

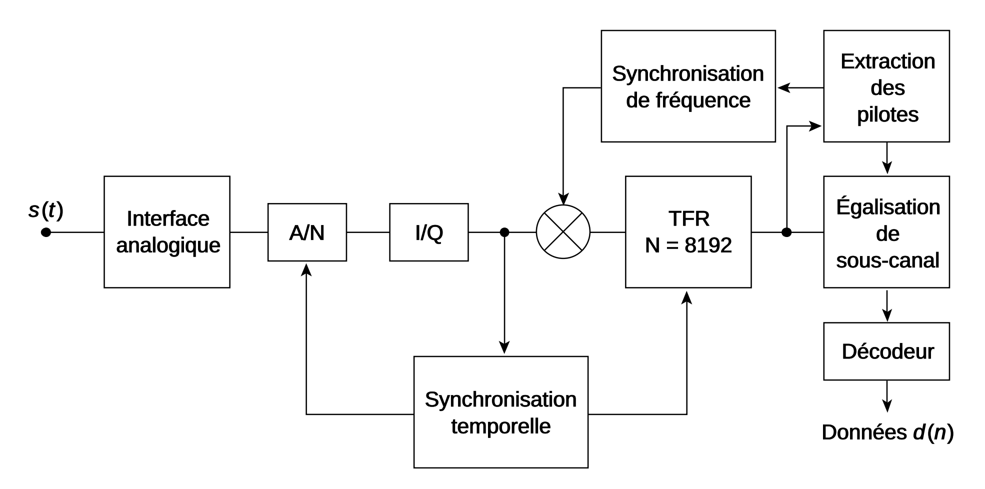


FIG. 15.15. *Récepteur de télévision numérique terrestre*

Les avantages de la technique OFDM résident dans sa faible sensibilité aux distorsions du canal, une relative immunité aux bruits impulsifs par l'effet de moyennage de la TFD, la possibilité d'éviter les brouilleurs à spectre étroit et l'ajustement du débit à la densité spectrale du bruit.

Par contre, cette technique nécessite des dispositifs de synchronisation délicats et elle est sensible aux non-linéarités. En particulier, il faut bien noter que le signal émis dans le canal étant la somme d'un grand nombre de signaux aléatoires de même distribution, il a une distribution d'amplitude gaussienne et un facteur de crête de 12 dB, ce qui est un inconvénient par rapport aux modulations à enveloppe constante, utilisées en radiocommunication par exemple.

Les intervalles de garde en temps et en fréquence, les signaux pilotes, les symboles de référence réduisent l'efficacité du système. Des approches basées sur le principe des bancs de filtres à décomposition et reconstitution presque parfaites, ce qui limite le recouvrement des sous-canaux aux voisins immédiats, permettent d'éviter ces pertes, au prix d'un supplément de traitement. Avec des égaliseurs adaptatifs dans chaque sous-canal, ils réduisent aussi les contraintes de synchronisation et ils devraient permettre d'approcher les limites des capacités de transmission théoriques [12].

BIBLIOGRAPHIE

[1] U. Mengali And A.N.d'Andrea, Synchronization Techniques for Digital Receivers, Plenum Press, 1997.

[2] H. Meyr, M. Moeneclay and S.A.Fechtel, Digital Communication Receivers : Synchronization, channel estimation and Signal Processing, Wiley, 1998.

[3] N. Moreau, Techniques de Compression des Signaux, Masson Ed., 1995.

[4] CCITT, Réseaux Numériques, Systèmes de Transmission et équipements de multiplexage, Tome III.3, Genève, 1985.

[5] ISO-CEI 11172, Technologies de l'information. Codage des images animées et du son, Genève, 1994.

[6] D. Lecomte, D. Cohen, P. de Bellefonds et J. Barda, Les Normes et les Standards du Multimedia, Dunod, 1999.

[7] C. Breining et al., " Acoustic Echo Control : an Application of Very-High-Order Adaptive Filters ", IEEE Signal Processing Magazine, Vol.16, N°4, July 1999, pp.42-69.

[8] F. Kretz – " Codage MIC-Differentiel à prédiction adaptative en télévision ", Annales des Télécommunications, N° 37, 1982.

[9] G. Sullivan and T. Wiegand, " Video Compression – From concepts to the H264/AVC standard ", IEEE Proc., vol.93, N° 1, January 2005, pp.18-31.

[10] T. de Couasnon, R. Monnier et J.B.Rault – " OFDM for Digital TV Broadcasting ", Signal Processing, Vol.39, 1994, pp.1-32.

[11] U. Ladebusch and C.A.Liss, " Terrestrial DVB – A broadcast technology for stationary portable and mobile use ", Proc. IEEE, vol.94, N° 1, Jan.2006, pp.183-193.

[12] L. Qin et M. Bellanger, " Equalisation issues in multicarrier transmission using filter banks ", Annales des Télécommunications, Vol.52, N° 1-2, 1997, pp.31-38.

EXERCICES

Éléments de réponse et indications

CHAPITRE 1

1.1 $I_L(t) = \frac{1}{2} + \frac{2}{\pi} \sum_{p=1}^{4} \frac{(-1)^{p+1}}{2p-1} \cos 2\pi(2p-1) \frac{t}{T}$.

1.2 $s(nT) = \sin(n\pi + \varphi) = (-1)^n \sin \varphi$.

La possibilité de reconstitution dépend de φ. $\left(\varphi = \frac{\varphi}{2} \text{ oui}; \varphi = 0 \text{ non}\right)$.

1.3 $H(fe/2) = \frac{2\sqrt{2}}{\pi}$ (0,92 dB).

1.4 $f_2 < f_e < 2f_1$.

1.5 $s(nT) = s_r(nT) + js_i(nT) = e^{j\frac{\pi}{2}n} \frac{\sin \frac{3\pi}{8} n}{\sin\left(\frac{\pi}{8} n\right)}$.

1.6 Valeur maximale de $s(n) = 8$;

$$s(n) = 0 \quad \text{pour} \quad \varphi_k = -2\pi \frac{k}{8} n + k\pi.$$

1.7 $fe = 2\text{MHz}$; $\Delta f = 1\ k\text{Hz}$.

1.8 $p(1) = \frac{1}{\pi} \frac{1}{\sqrt{A^2 - s^2}}$; $r(\tau) = 2(f_2 - f_1) \frac{\sin \pi(f_2 - f_1)\tau}{\pi(f_2 - f_1)\tau} \cos \pi(f_2 + f_1)\tau$.

1.9 Partie périodique ; coefficient de Fourier : $C_n = \frac{p \sin \frac{\pi}{2} n}{\pi n}$.

Partie non périodique ; spectre : $S_2(f) = p(1-p)T \frac{1 - \cos \pi f T}{\pi^2 f^2 T^2}$.

1.**10** Rapport signal à bruit dans la bande 300 – 500 Hz = 75 dB ($fe = 16\ k$Hz ; gain 3 dB).

1.**11** Distorsion de quantification : raie à $\frac{3}{8} fe$; puissance : $0{,}0195^2$.

1.**12** Si la caractéristique est centrée : $a_1 = 0$ pour $0 \leqslant |\alpha| \leqslant \frac{1}{2}$; $a_1 = \frac{4q}{\pi}\sqrt{1 - \frac{1}{4\alpha^2}}$ pour $0 \leqslant |\alpha| \leqslant 1$. Centrage à $\frac{q}{2}$: $a_1 = \frac{2q}{\pi}$ pour $0 \leqslant |\alpha| \leqslant 1$.

1.**13** Sans écrêtage (facteur de crête) : 10 bits ; avec écrêtage à 1 % = 9 bits.

1.**14** En codage linéaire $(S/B)_{max} = 50$ dB ; en non linéaire il varie de 35 à 38 dB quand le signal varie de – 36 dB à 0 dB.

1.**15** Valeurs optimales : $x_0 = 0$; $x_1 = 0{,}9816$; $y_1 = 0{,}4528$; $y_1 = 1{,}510$.

CHAPITRE 2

2.**1** La TFD de la seconde suite est liée à la première par :

$$X'(k) = e^{-jk\frac{\pi}{4}} X(k)$$

2.**2** Multiplications réelles : $M_R = 28$; Additions réelles : $A_R = 84$.

2.**3** Les petites différences proviennent des repliements dans l'échantillonnage et décroissent quand N croît.

2.**4** Nombres de multiplications complexes : 160, 96, 72. Additions : 384.

2.**5** Puissance maximale de bruit sur une sortie : $28 . q^2/12$. Avec quantification à 8 bits des coefficients : $|\varepsilon(i, k)| \leqslant 0{,}003$.

2.**6** Récurrence : $X_0 = x(N-1)$; $X_m = x(N-1-m) + WX_{m-1}$ pour

$$1 \leqslant m \leqslant N-1$$

Il faut N – 1 multiplications complexes.

2.**7** Bruit d'arrondi total : $N\frac{q^2}{12} + Nq^2$; dégradation de rapport signal à bruit : $\Delta SB = 11{,}5$ dB (Bruit d'entrée $q^2/12$).

2.**8** Enregistrement : 20000 échantillons ; mémoire 160 *k*bits temps de cycle pour une multiplication : 1µs.

2.**9** Les fenêtres en cosinus, Hamming et Blackman affaiblissent les lobes secondaires, mais ne permettent plus de détecter la présence de la composante faible.

2.**10** Sur la figure 2.13 les multiplieurs sont utilisés à 50 %. La pleine capacité est obtenue en doublant la mémoire d'entrée avec lecture alternée.

2.**11** Mémoires : 120, 30 et 6 nombres dans les 3 étapes respectivement. Trois multiplieurs complexes sont nécessaires dans deux étapes ; leur rendement peut être porté à 100 % par des mémoires tampon.

CHAPITRE 3

3.**1** On vérifie que les produits $I_3 \times A$ et $A \times I_3$ sont différents.

3.**2** On vérifie les relations 3.3, 3.4, 3.5, 3.6.

3.**3** Nombres de multiplieurs réels en bases 2, 4 et 8 : 384, 284 et 246.

3.**4** La TFD d'ordre 12 demande 20 multiplications complexes.

3.**5** Utiliser les relations (3.18) et (3.21) pour obtenir les 2 factorisations.

3.**6** Le calcul basé sur une TFD complexe d'ordre 8 conduit à 24 multiplications réelles, la transformée impaire à 26.

3.**7** Avec cette méthode les opérations de Δ_{12} disparaissent, ce qui réduit à 16 le nombre de multiplications complexes.

3.**8** Matrices de la transformation :

$$T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 4 & 16 & 13 \\ 1 & 16 & 1 & 16 \\ 1 & 13 & 16 & 4 \end{bmatrix} ; \; T^{-1} = 13 \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 13 & 16 & 4 \\ 1 & 16 & 1 & 16 \\ 1 & 4 & 16 & 13 \end{bmatrix}$$

CHAPITRE 4

4.**1** Réponse à la suite a^n : $y(n) = a^{n-3} \dfrac{1 - a^{9-n}}{1-a}$ pour $5 \leqslant n \leqslant 8$.

4.**2** Procéder par dérivation et développement en série et intégration

$$\ln(z-a) = -\sum_{n=1}^{\infty} \frac{1}{na^n} Z^n ; \quad \frac{Z^{-1}}{(1-aZ^{-1})(1-bZ^{-1})} = \frac{1}{a-b} \sum_{n=0}^{\infty} \left(\frac{1}{b^n} - \frac{1}{a^n} \right) Z^{-n}.$$

avec $|a| < 1; \quad |b| < 1$

4.**3** $H(Z) = \dfrac{1}{(1 - re^{j\theta}Z^{-1})(1 - re^{-j\theta}Z^{-1})}$.

4.**4** Puissance en sortie : 21 ; $H(\omega) = 4{,}41 - 1{,}536 \cos \omega + 0{,}46 \cos 2\omega$.

4.**5** Réponse du système :

$$y(n) = \frac{e^{jn\omega}}{1 - e^{-j\omega} + 0{,}8\, e^{-2j\omega}} + r^{n-1} \left[(a - 0{,}8b) \frac{r \sin(n+1)\theta}{\sin\theta} - 0{,}8a \frac{\sin(n\theta)}{\sin\theta} \right]$$

CHAPITRE 5

5.**1** La réponse s'annule pour $f = 0{,}288$; 0,347 ; 0,408 ; 0,469 ondulation maximale : 0,08. Zéros de H(Z) : 0,606 ; 1,651 ; $0{,}4292 \pm j0{,}464$; $1{,}073 \pm j1{,}161$.

5.**2** Coefficients : – 0,012 ; 0 ; 0,042 ; 0 ; – 0,093 ; 0 ; 0,314 ; 0,5. Zéros de H(Z) : 0,4816 ; 2,076 ; $0{,}3764 \pm j0{,}368$; $1{,}3583 \pm j1{,}328$ ondulation maximale : 0,03.

5.**3** $\delta = 0{,}017$, valeur inférieure aux valeurs ci-dessus.

5.**4** Avec *fs*/2 en sortie, les nombres de mémoires et multiplications sont divisés par deux, par entrelacement (voir § 10.5).

5.**5** Dans le plan complexe H(Z) tourne de π et de $\pm \frac{\pi}{2}$, ce qui donne un passe-haut et un passe-bande.

5.**6** Coefficients : N = 27 ; précision des calculs : $b_c = 12$ bits ; $b_i = 20$ bits.

5.**7** Coefficients de la fonction erreur : – 0,0065 ; 0,0034 ; – 0,0015 ; – 0,0019
$e(0{,}1925) = e^{-j2\pi 8 f_0}(-0{,}0028)$

$$b_c \simeq 1 + \frac{1}{2}\log_2\left(\frac{f_e}{2\Delta f}\right) + \log_2\left(\frac{1}{\min\{\delta_1, \delta_2\}}\right).$$

CHAPITRE 6

6.**1** Suivre le développement du paragraphe (6.1). La différence entre le retard et le temps de groupe du filtre illustre la non-linéarité de la réponse en phase.

6.**2** Réponse à l'échelon unité : $y(n) = \frac{1}{1{,}8}\left(1 - (-0{,}8)^{n+1}\right) + (-0{,}8)^{n+1}\, y(-1)$.

6.**3** Pôles : $P = 0{,}78 \pm j0{,}438$. Les zéros n'ajoutent pas de multiplications dans le circuit.

6.**4** $H_m = 85$; $\cos \omega_0 = 0{,}808$; $\|H\|_2 = 8{,}53$.
Avec des zéros à $3f_e/8$: $\|H\|_2 = 25{,}8$.
Fréquence des auto-oscillations voisine de $f_e/10$; amplitude $\simeq 42q$. Auto-oscillations de forte amplitude possible car (6.56) non-vérifiée.

6.**5** $b_c \geqslant 13$ bits ; déplacement de la pointe d'affaiblissement : $\mathrm{d}f_i \leqslant 2{,}210^{-5} f_e$.

$$H(Z) = \frac{0{,}796 - 1{,}42\,Z^{-1} + Z^{-2}}{1 - 1{,}42\,Z^{-1} + 0{,}796\,Z^{-2}}\ ; \ \tau_g(\omega) \text{ calculé par (6.45)}$$

6.**6** Réalisation possible avec 3 multiplications.

CHAPITRE 7

7.1 Cellule du 1er ordre : $|H(\omega)|^2 = \dfrac{1{,}49 + 1{,}4 \cos \omega}{1{,}81 - 1{,}8 \cos \omega}$

$$\varphi(\omega) = \text{Arctg}.\ \frac{1{,}6 \sin \omega}{1{,}63 - 0{,}2 \cos \omega}\ ;\ \tau(\omega) = \frac{1{,}6(0{,}37 \cos \omega - 0{,}2)}{(0{,}37 \cos \omega - 0{,}2)^2 + 2{,}66 \sin^2 \omega}$$

7.2 Fréquences caractéristiques : 0,162 ; 0,231 ; 0,538 ; 0,736.

7.3 Fonction de transfert : $H(Z) = 0{,}094\ \dfrac{(1 + Z^{-1})^4}{(1 + 0{,}039 Z^{-2})\,(1 + 0{,}447 Z^{-2})}$.

7.4 Transformation : $Z^{-1} = -\dfrac{Z^{-1} - \alpha}{1 - \alpha Z^{-1}}\ ;\ \alpha = \dfrac{\sin \pi(f_1 - f'_1)}{\sin \pi(f_1 + f'_1)}$

avec $f_1 = 0{,}1725\ ;\quad f'_1 = 0{,}1\ ;\quad \alpha = 0{,}3.$

7.5 Facteurs d'échelle : $a_0^0 = 2^{-6}$; $a_0^1 = 2^{+2}$; $a_0^2 = 2$.
Pour $H(j) = 1$, il vient : $a_0 = 0{,}515$ et $a_0^3 = 4{,}12$; Précision : $b_i = 16$ bits.

7.6 Nombre de bits des coefficients : $b_c \simeq 12$ bits. Optimum obtenu par recherche systématique autour de l'arrondi. Le pôle critique $0{,}9235 \pm j0{,}189$ ne permet pas de ramener b_c à 11 bits.

7.7 Le filtre du paragraphe 7.2.2 peut avoir des auto-oscillations d'amplitude inférieure à $3q$ et de fréquence voisine de fe/5. De même pour le filtre de la figure 7.20.

7.8 Le filtre RII nécessite 7 multiplications et 4 mémoires alors que le RIF équivalent demande 8 multiplications et 16 mémoires.

7.9 Fonction de transfert :

$$H(Z) = 0{,}0625\ \frac{1 - 1{,}165 Z^{-1} + Z^{-2}}{1 - 1{,}404 Z^{-1} + 0{,}84 Z^{-2}}\ \frac{1 - 0{,}198 Z^{-1} + Z^{-2}}{1 - 1{,}238 Z^{-1} + 0{,}455 Z^{-2}}$$

$$f_1 = 4\,832 \text{ Hz}\ ;\quad f_2 = 7\,495 \text{ Hz}\ ;\quad \Delta f_1 = 4 \text{ Hz}\ ;\quad \Delta f_2 = -3 \text{ Hz}.$$

Facteurs d'échelle : $a_0^0 = 2^{-3}$; $a_0^1 = 2^{-1}$; $a_0^2 = 1$.

7.10 Ordre théorique : N = 5,19 ; pour N = 6. δ_1 devient très faible. Nombre de bits des coefficients : $b_c \approx 11$ bits. Différence entre données internes et entrée : 7 bits.

CHAPITRE 8

8.1

$$S = \frac{1}{z+2}\begin{bmatrix} z & 2 \\ 2 & z \end{bmatrix};\quad t = \begin{bmatrix} 1 - z/2 & z/2 \\ -z/2 & 1 + z/2 \end{bmatrix}$$

Pour des circuits LC, prendre $z = Lp + \dfrac{1}{Lp}$ ou $z = \dfrac{LC}{Lp + \dfrac{1}{Cp}}$.

8.**2** Le diagramme est celui de la figure 8.3 avec N = 6 et $Y_6 = 0$. Pour

$$f_e = 40\ k\text{Hz}\,; \quad a_1 = a_4 = 0{,}205\,; \quad a_2 = a_3 = 0{,}085\,;$$

les coefficients sont multipliés par 4 pour $f_e = 10$ kHz.

8.**3** Le circuit de la figure 8.2 peut être utilisé. Les produits $sL_2 \cdot \frac{1}{sL_3}$ et $sL_2 \cdot \frac{1}{sL_1}$ sont réalisés par des connections directes de l'entrée de la branche centrale aux deux additionneurs. Valeurs des coefficients : $a_1 = 5{,}5 10^{-4}$; $a_2 = 2{,}6 10^{-4}$; $a_3 = 2{,}4 10^{-4}$; $a_4 = 4{,}6 10^{-4}$; $L_2/L_1 = 0{,}097$; $L_2/L_3 + L_2/L_1 = 0{,}32$.

8.**4** Coefficients :

$$\alpha_1 = 0{,}4425\,;\ \alpha_3 = 0{,}1856\,;\ \alpha_5 = 0{,}1793\,;\ \alpha_7 = 0{,}7359$$
$$\beta_2 = 0{,}2255\,;\ B_4 = 0{,}1781\,;\ \beta_6 = 0{,}1944\,;\ \alpha_7' = 0{,}7169$$

Avec $b_c = 5$ bits le filtre d'onde a moins d'ondulations.

8.**5** Zéros du filtre en treillis : 0,6605 ; $0{,}6647 \pm j0{,}5020$ après arrondi à 5 bits des k_i : 0,6661 ; $0{,}6377 \pm j0{,}5002$.

CHAPITRE 9

9.**1** $$X(f) = \frac{1 - a\cos 2\pi f}{1 + a^2 - 2a\cos 2\pi f} + j\,\frac{-a\sin 2\pi f}{1 + a^2 - 2a\cos 2\pi f}.$$

9.**2** Calculer $X_I(\omega)$ par transformation de Hilbert ; ou écrire

$$X_R(\omega) = \frac{1}{2}\left[\frac{1}{1-p} + \frac{1}{1-\bar{p}}\right].$$

9.**3** Les séquences $x_R(n)$ et $x_I(n)$ ont leurs termes non nuls entrelacés. L'opération faite par $x(n)$ est un filtrage analytique ; $y(n) = \frac{1}{2}\, e^{-jn}\,\frac{\pi}{5}$.

9.**4** Nombre de bits des coefficients : $b_c \simeq 2 + \frac{1}{2}\log_2\left(\frac{f_e}{2\Delta f}\right) + \log_2\left(\frac{1}{\delta}\right)$.

9.**5** Ordre du déphaseur : $N \simeq \log\left(\frac{\pi}{\varepsilon}\right)\log\left(\frac{f_e}{f_1}\,\frac{f_e}{f_2}\right)$;

coefficients : $b_c \approx \log_2\left(\frac{\pi}{\varepsilon}\right) + \log_2\left(\frac{f_e}{f_1}\right) + \log_2\left(\frac{f_e}{f_2}\right)$.

Pour l'exemple du paragraphe 9.4 : N = 4,97 ; $b_c \simeq 14$ bits;

9.**6** Fonctions de transfert :

$$H_m(Z) = 1 - 2Z^{-1} + 2Z^{-2} - Z^{-3} + 0{,}25Z^{-4}$$
$$H_L(Z) = 0{,}5 - 1{,}5Z^{-1} + 2{,}25Z^{-2} - 1{,}5Z^{-1} + 0{,}5Z^{-4}$$
$$H_M(Z) = 0{,}25 - Z^{-1} + 2Z^{-2} - 2Z^{-3} + Z^{-4}$$

9.**7** Réponse en fréquence du filtre

$$H(f) = e^{-j2\pi 3f}\,[0{,}5 + 0{,}5902 \cos 2\pi f - 0{,}1012 \cos 6\pi f]$$

$$H(0) = H\left(\frac{f_e}{8}\right) = 0{,}989\,; \quad \delta_1 = \delta_2 = 0{,}011\,; \quad \Delta f = \frac{f_e}{4}$$

Sortie du modulateur IQ.

$$y(n) = 0{,}4945\, e^{-j(n-5)\frac{\pi}{4}} + 0{,}0055\, e^{j(n-5)\frac{\pi}{4}}$$

Après sous-échantillonnage, en posant $n = 2p + 1$

$$y(p) = -\,0{,}4945 e^{-jp\frac{\pi}{2}} - 0{,}0055 e^{jp\frac{\pi}{2}}.$$

Deux composantes à $\frac{f_e}{4}$ et $-\frac{f_e}{4} = 3\frac{f_e}{4}$.

9.**8** Bande passante : [0 ; 0,25] ; Ondulation : 4.10^{-2} ; retard : 2T

CHAPITRE 10

10.**1** Nombre de bits des coefficients b_c = 1, 2, 5, 6, 9, 10, 10, 11, 14. Pour le filtre demi-bande : $b_c \approx 2 + \mathrm{lb}\,(1/\delta_m - \delta_0)$.

10.**2** Filtres dans la suite de trois filtres : $\Delta f = 0{,}4$ avec M = 2 ; $\Delta f = 0{,}15$ avec M = 3 ; $\Delta f = 0{,}025$ avec M = 8. Bruit de calcul en sortie d'un filtre demi-bande : $2\mathrm{M}\,\frac{q^2}{12}$. Après les 3 filtres, $P_B = 20\,\frac{q^2}{12}$.

10.**3** La fonction peut se réaliser avec un filtre demi-bande (M = 3) et un filtre passe-bas à 54 coefficients, d'où une capacité de calcul de 264 *k*mult/s. Une réalisation directe à 100 coefficients conduit à 400 *k*mult/s.

10.**4** La TFD impaire correspond à un décalage en fréquence de $f_e/2N$.

10.**5** Fonctions du réseau polyphasé :

$$D(Z)\,(1 - 0{,}1354\,Z^{-1} + 0{,}069\,Z^{-2})\,(1 + 0{,}98\,Z^{-1} + 0{,}51\,Z^{-2})$$
$$N_1(Z) = 1 + 7{,}806\,Z^{-1} + 9{,}718\,Z^{-2} + 3{,}773\,Z^{-3} + 0{,}1883\,Z^{-4}$$
$$N_2(Z) = 3{,}713\,(1 + 2{,}908\,Z^{-1} + 2{,}035\,Z^{-2} + 0{,}317\,Z^{-3}).$$

Le schéma est comparable à celui de la figure 10.1 ; cadence des mult. : $8f_e$.

CHAPITRE 11

11.**1** Réponse en fréquence : $H(f) = e^{-j2\pi f 2{,}5}[0{,}904\cos\pi f + 0{,}234\cos 3\pi f - 0{,}1\cos 5\pi f]$

Signal de sortie : $y(n) = 0{,}963\cos(n-2{,}5)\frac{\pi}{4}$

En sortie des filtres d'analyse, après sous-échantillonnage :

$$u_1(n) = 0{,}963\cos(n-2{,}5)\frac{\pi}{4} + 0{,}963\cos(n-2{,}5)\frac{3\pi}{4}$$

$$u_2(n) = 0{,}037\cos(n-2{,}5)\frac{\pi}{4} + 0{,}037\cos(n-2{,}5)\frac{3\pi}{4}$$

Fonction de transfert totale :

$$T(2) = Z^{-1}[-0{,}023 + 0{,}121\,Z^{-2} + 0{,}882\,Z^{-4} + 0{,}121\,Z^{-b} - 0{,}023\,Z^{-8}]$$

Signal reconstitué : $x'(n) = 0{,}929\cos(n-5)\frac{\pi}{4}$

11.**2** Fonctions de transfert des deux facteurs

$$H_0(Z) = \frac{1}{4}(1+Z^{-1})^3\ ;\ H_1(Z) = \frac{1}{4}(-1-32^{-1}+32^{-2}+Z^{-3})$$

Facteur d'amplification : $\frac{20}{16} = 1{,}25$

11.**3** Reprendre la procédure du paragraphe 11.3 sans imposer le zéro double au point – 1 pour $H_1(-Z)$

11.**4** Signaux de sortie : $y_1(n) = 0{,}951\cos(n-4)\frac{\pi}{4}$

$$y_2(n) = 0{,}15\cos(n-3)\frac{\pi}{4}$$

Le sous-échantillonnage introduit les composantes images à la fréquence 3/8.

On vérifie que les composantes images s'annulent à la reconstitution.

11.**5** L'erreur de reconstitution est bornée par le pas de quantification multiplié par le double de la somme des valeurs absolues des coefficients.

CHAPITRE 13

13.**1** Fonction AC : $r(0) = 1$; $r(1) = 0{,}707$; $r(2) = 0$

Valeurs propres de R_3 : N = [2, 1, 0]

13.**2** Puissance de sortie :

$$P_s = (1 + a_1^2 + a_2^2)V_b^2 + |1 - a_1 e^{-j\omega} - a_2^{-j\omega}|^2$$

En annulant les dérivées, il vient :

$$a_1 = 2\cos\omega \; \frac{\sin^2\omega + \sigma_b^2/2}{\sin^2\omega + \sigma_b^2(2 + \sigma_b^2)} \; ; \; a_2 = -1\left[1 - V_b^2 \frac{1 + \sigma_b^{22} + 2\cos^2\omega}{(1 + \sigma_b^{22})^2 - \cos^2\omega}\right]$$

13.**3** Fonction AC : $r(0) = 1$; $r(1) \; \frac{1}{2}(\cos\frac{\pi}{4} + \cos\frac{\pi}{3}) \; r(2) = \frac{1}{4}\cos\frac{2\pi}{3}$

On vérifie que les zéros du prédicteur se situent entre les points d'affixes $e^{j\frac{\pi}{4}}$ et $e^{j\frac{\pi}{3}}$

13.**4** On vérifie que les racines des polynomes obtenues sont sur le cercle unité et vérifient le principe d'alternance.

CHAPITRE 14

14.**1** Constante de temps $\tau = \frac{1}{\delta}$; il faut 23 échantillons pour que $y(n)$ approche m à 1 % en moyenne. Après la transition, l'erreur quadratique est donnée par (11.39). Les estimateurs récursifs et non récursifs sont équivalents pour $n \simeq \frac{2}{\delta}$.

14.**2** Le filtre de prédiction présente un affaiblissement infini à la fréquence fe/8 ; d'où :

$a_1 = \sqrt{2}$; $\quad a_2 = -1$.

avec un bruit σ^2 il vient :

$$a_1 = \sqrt{2}\;\frac{1 + 2\sigma^2}{1 + 8\sigma^2 + 8\sigma^4} \simeq \sqrt{2}(1 - 6\sigma^2); \quad a_2 = -\frac{1}{1 + 8\sigma^2 + 8\sigma^4} \approx -(1 - 8\sigma^2)$$

14.**3** Fonction d'autocorrélation du signal

$$r_0 = P_x = \frac{1}{3} \; ; \quad r_1 = \frac{1}{6} \; ; \quad r_2 = \frac{1}{12}$$

Valeur optimale des coefficients : $H_{opt}^t = [2 ; -1; \; 0]$

Avec bruit : $H_{opt}^t = [1{,}35 - 0{,}50 - 0{,}07]$

Facteur d'amplification du bruit $\|H_{opt}\|_2^2 = 2{,}09$

De la réponse totale (Canal + égaliseur), on déduit la puissance de l'interférence résiduelle.

Valeur propre minimale $\lambda_{min} = 0{,}235$. On vérifie que la constante de temps en simulation correspond à l'estimation avec λ_{min}.

14.**4** Donner l'expression de l'erreur de sortie et rechercher les valeurs des coefficients qui minimisent sa puissance.

On pourra commencer par remplacer, dans la relation d'entrée-sortie du filtre $H(Z)$, la sortie $\tilde{y}(n)$ par la référence $y(n)$ et calculer les valeurs optimales des coefficients dans ce cas.

INDEX ALPHABÉTIQUE

E

F

G

H

I

K

L

BIBLIOGRAPHIE

Aux éditions Masson

TÉLÉCOMMUNICATIONS SPATIALES, par des ingénieurs du CNES et du CNET.
Tome 1. – Bases théoriques, 1982, 432 pages.
Tome 2. – Secteur spatial. 1983, 400 pages.
Tome 3. – Secteur terrien. Systèmes de télécommunications par satellites. 1983, 468 pages.

STÉRÉOPHONIE. Cours de relief sonore théorique et appliqué, par R. CONDAMINES. 1978, 320 pages.

DÉCISIONS EN TRAITEMENT DU SIGNAL, par P.-Y. ARQUÈS. 1982, 2e édition, 288 pages.

LES RÉSEAUX PENSANTS. Télécommunications et société, sous la direction de A. GIRAUD, J.-L. MISSIKA et D. WOLTON. 1978, 206 pages (*épuisé*).

LES FILTRES NUMÉRIQUES. Analyse et synthèse des filtres unidimensionnels, par R. BOITE et H. LEICH. 1990, 3e édition révisée et augmentée, 432 pages.

TRAITEMENT NUMÉRIQUE DU SIGNAL. Théorie et pratique, par M. BELLANGER. 1990, 6e édition révisée et augmentée, 480 pages.

FONCTIONS ALÉATOIRES, par A. BLANC-LAPIERRE et B. PICINBONO. 1981, 440 pages.

PSYCHOACOUSTIQUE. L'oreille récepteur d'information, par E. ZWICKER et R. FELDKELLER. Traduit de l'allemand par C. Sorin. 1981, 248 pages.

GENÈSE ET CROISSANCE DES TÉLÉCOMMUNICATIONS, par L.-J. LIBOIS. 1983, 432 pages.

LE VIDÉOTEX. Contribution aux débats sur la télématique, coordonné par Cl. ANCELIN et M. MARCHAND. 1984, 256 pages.

ÉCOULEMENT DU TRAFIC DANS LES AUTOCOMMUTATEURS, par G. HÉBUTERNE. 1985, 264 pages.

L'EUROPE DES POSTES ET DES TÉLÉCOMMUNICATIONS, par Cl. LABARRÈRE. 1985, 256 pages.

THÉORIE STRUCTURALE DE LA COMMUNICATION ET SOCIÉTÉ, par A. A. MOLES. 1986, 296 pages.

TRAITEMENT DU SIGNAL PAR ONDES ÉLASTIQUES DE SURFACE, par M. FELDMAN et J. HÉNAFF. 1986, 400 pages.

THÉORIE DE L'INFORMATION OU ANALYSE DIACRITIQUE DES SYSTÈMES, par J. OSWALD. 1986. 448 pages.

LES VIDÉODISQUES, par G. BROUSSAUD, 1986, 216 pages.

LES PARADIS INFORMATIONNELS. Du Minitel aux services de communication du futur, par M. MARCHAND et le SPES. 1987, 256 pages.

LES MODEMS POUR TRANSMISSION DE DONNÉES, par M. STEIN, 1987, 384 pages.

SYSTÈMES ET RÉSEAUX DE TÉLÉCOMMUNICATION EN RÉGIME STOCHASTIQUE, par G. DOYON. 1989, 704 pages.

PRINCIPES DE TRAITEMENT DES SIGNAUX RADAR ET SONAR, par R. LECHEVALIER. 1989, 280 pages.

CIRCUITS INTÉGRÉS EN ARSÉNIURE DE GALLIUM. Physique, technologie et règles de conception, par R. CASTAGNÉ, J.-P. DUCHEMIN, M. GLOANEC et C. RUMELHARD. 1989, 616 pages.

ANALYSE DES SIGNAUX ET FILTRAGE NUMÉRIQUE ADAPTATIF, par M. BELLANGER. 1989, 404 pages.

LA PAROLE ET SON TRAITEMENT AUTOMATIQUE, par CALLIOPE. 1989, 736 pages.

LE RNIS. Techniques et atouts, par J. DICENET. 1990, 384 pages.

Aux éditions Dunod

TÉLÉCOMMUNICATIONS PAR FAISCEAU HERTZIEN, par M. MATHIEU. 1980, 2e tirage, 334 pages (*épuisé*).

TÉLÉCOMMUNICATIONS : OBJECTIF 2000, sous la dir. de A. GLOWINSKI. 1981, 2e tirage, 300 pages (*épuisé*).

ÉLECTROMAGNÉTISME CLASSIQUE DANS LA MATIÈRE, par Ch. VASSALLO. 1980, 272 pages.

PRINCIPES DES COMMUNICATIONS NUMÉRIQUES, par A.-J. VITERBI et J.-K. OMURA. Traduit de l'anglais par G. Batail. 1982, 232 pages.

PROPAGATION DES ONDES RADIOÉLECTRIQUES DANS L'ENVIRONNEMENT TERRESTRE, par L. BOITHIAS. 1984, 2e tirage, 328 pages.

PROGRAMMATION MATHÉMATIQUE. Théorie et algorithmes, par M. MINOUX.
Tome 1. – 1983, 328 pages.
Tome 2. – 1983, 272 pages.

SYSTÈMES DE TÉLÉCOMMUNICATIONS. Bases de transmission, par P.-G. FONTOLLIET. 1984, 2e tirage, 528 pages.

ÉLÉMENTS DE COMMUNICATIONS NUMÉRIQUES. Transmission sur fréquence porteuse, par J.-C. BIC, D. DUPONTEIL et J.-C. IMBEAUX.
Tome 1. – 1986, 363 pages.
Tome 2. – 1986, 328 pages.

TÉLÉINFORMATIQUE. Transport et traitement de l'information dans les réseaux et systèmes téléinformatiques et télématiques, par C. MACCHI, J.-F. GUILBERT et 13 co-auteurs. 1987, nouvelle édition entièrement revue et augmentée, 934 pages.

LES SYSTÈMES DE TÉLÉVISION EN ONDES MÉTRIQUES ET DÉCIMÉTRIQUES, par L. GOUSSOT. 1987, 376 pages.

COMPATIBILITÉ ÉLECTROMAGNÉTIQUE, sous la coordination de P. DEGAUQUE ET J. HAMELIN. A paraître.

LES FAISCEAUX HERTZIENS ANALOGIQUES ET NUMÉRIQUES, par E. FERNANDEZ et M. MATHIEU. A paraître.

Aux éditions Eyrolles

DE LA LOGIQUE CÂBLÉE AUX MICROPROCESSEURS, par J.-M. BERNARD et J. HUGON.
Tome 1. – Circuits combinatoires et séquentiels fondamentaux, avec la collaboration de R. LE CORVEC. 1983, 6e tirage, 232 pages (*épuisé*).
Tome 2. – Applications directes des circuits fondamentaux. 1983, 6e tirage, 148 pages (*épuisé*).
Tome 3. – Méthodes de conception des systèmes. 1986, 7e tirage, 164 pages.
Tome 4. – Application des méthodes de synthèse. 1987, 8e tirage, 272 pages.

LA COMMUNICATION ÉLECTRONIQUE, par CRINSEC.
Tome 1. – Structure des systèmes spatiaux et temporels. 1984, 3e tirage, 456 pages (*épuisé*).
Tome 2. – Logiciel. Mise en œuvre des systèmes. 1984, 3e tirage, 512 pages.

OPTIQUE ET TÉLÉCOMMUNICATIONS. Transmission et traitement optiques de l'information, par A. COZANNET, J. FLEURET, H. MAITRE et M. ROUSSEAU. 1983, 2e tirage, 512 pages (*épuisé*).

THÉORIE DES RÉSEAUX ET SYSTÈMES LINÉAIRES, par M. FELDMANN. 1987, 2e édition entièrement revue et augmentée, 424 pages.

RADARMÉTÉOROLOGIE. Télédétection active de l'atmosphère, par H. SAUVAGEOT. 1982, 304 pages.

PRATIQUE DES CIRCUITS LOGIQUES, par J.-M. BERNARD et J. HUGON. 1987, 3e tirage, 472 pages.

THÉORIE DES GUIDES D'ONDES ÉLECTROMAGNÉTIQUES, par Ch. VASSALLO.
Tome 1. – 1985, 504 pages.
Tome 2. – 1985, 700 pages.

CONCEPTION DES CIRCUITS INTÉGRÉS MOS. Éléments de base et méthodologies, par M. CAND, E. DEMOULIN, J.-L. LARDY et P. SENN. 1986, 432 pages.

THÉORIE DES RÉSEAUX ET SYSTÈMES LINÉAIRES, par M. FELDMANN. 1987, 2e édition entièrement revue et augmentée, 424 pages.

CONCEPTION STRUCTURÉE DES SYSTÈMES LOGIQUES, par J.-M. BERNARD. 1987, 386 pages.

A la Documentation française

LES TÉLÉCOMMUNICATIONS FRANÇAISES. QUEL STATUT POUR QUELLE ENTREPRISE ?, par Geneviève BONNETBLANC. 1985, 240 pages.

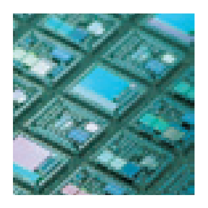
LA COMMUNICATION AU QUOTIDIEN. De la tradition et du changement à l'aube de la Vidéocommunication par Josiane JOUET avec la collaboration de Nicole CELLE. 1985, 240 pages.

050162 - (I) - (1,5) - OSB 80°- CP2- ABS

Imprimerie Nouvelle
45800 Saint-Jean-de-Braye
N°d'Imprimeur : 418537N
Dépôt légal : août 2006

Imprimé en France

SCIENCES SUP



8[e] édition

Maurice Bellanger

TRAITEMENT NUMÉRIQUE DU SIGNAL

Théorie et pratique

Le traitement du signal se généralisant avec l'électronique, le domaine d'application de cette discipline à part entière est de plus en plus vaste.

Cet ouvrage est donc l'outil indispensable des étudiants en Master et des élèves des écoles d'ingénieurs qui désirent connaître les principales techniques de traitement numérique du signal.

Dans le domaine du traitement du signal, les techniques numériques apportent aujourd'hui des possibilités prodigieuses : une conception rigoureuse des systèmes, une reproductibilité aisée des équipements, une fiabilité des caractéristiques exploitées et une facilité de supervision. Cependant, ces techniques présentent un certain degré d'abstraction. Cet ouvrage permet de faciliter l'accès aux techniques numériques en reliant la théorie et la pratique.

Cette 8[e] édition entièrement révisée présente des chapitres complémentaires afin de montrer les connexions qui existent entre les filtres QMF et les ondelettes. L'analyse et la modélisation des signaux qui s'intègrent de plus en plus dans les systèmes sont également étudiées.

Des applications pratiques et des exercices corrigés sont proposés à la fin de chaque chapitre afin que le lecteur puisse valider de manière concrète ses acquis.

MAURICE BELLANGER

est professeur au Conservatoire National des Arts et Métiers à Paris et membre de l'Académie des technologies.









ISBN 2 10 050162 3

www.dunod.com